An on-chip photonic deep neural network for image classification

https://doi.org/10.1038/s41586-022-04714-0

Received: 10 June 2021

Accepted: 1 April 2022

Published online: 01 June 2022

Check for updates

Farshid Ashtiani[1], Alexander J. Geers[1] & Firooz Aflatouni[1]✉

Deep neural networks with applications from computer vision to medical diagnosis[1–5] are commonly implemented using clock-based processors[6–14], in which computation speed is mainly limited by the clock frequency and the memory access time. In the optical domain, despite advances in photonic computation[15–17], the lack of scalable on-chip optical non-linearity and the loss of photonic devices limit the scalability of optical deep networks. Here we report an integrated end-to-end photonic deep neural network (PDNN) that performs sub-nanosecond image classification through direct processing of the optical waves impinging on the on-chip pixel array as they propagate through layers of neurons. In each neuron, linear computation is performed optically and the non-linear activation function is realized opto-electronically, allowing a classification time of under 570 ps, which is comparable with a single clock cycle of state-of-the-art digital platforms. A uniformly distributed supply light provides the same per-neuron optical output range, allowing scalability to large-scale PDNNs. Two-class and four-class classification of handwritten letters with accuracies higher than 93.8% and 89.8%, respectively, is demonstrated. Direct, clock-less processing of optical data eliminates analogue-to-digital conversion and the requirement for a large memory module, allowing faster and more energy efficient neural networks for the next generations of deep learning systems.

Inspired by the distributed data processing in the human brain, deep neural networks are designed to process the input data using interconnected layers of neurons (nodes), which can be trained using a set of training data to learn a specific task. Once trained, the network can be used to perform the same task on a new set of data with high accuracies. Figure 1a shows a general architecture of a deep neural network, in which the input data are first arranged and then processed using the neurons of the first layer, followed by the intermediate (hidden) layers. The classification result appears at the output of the last (output) layer. Each neuron in the network generates an output by passing the weighted sum of its inputs through a non-linear activation function (Fig. 1b).

Deep neural networks are usually implemented using digital-clock-based platforms, such as graphics processing units (GPUs)[13,14] or application-specific integrated circuits (ASICs)[18,19]. GPUs are highly reconfigurable processors that are capable of performing a large number of computations in parallel, yet their computation time is mainly limited by the clock frequency (mostly less than 3 GHz for state-of-the-art GPUs) and the memory access time[20]. Implementation of deep networks using ASICs can provide one to two orders of magnitude improvement in terms of performance per unit energy consumption compared with GPUs[21]. However, they generally face similar challenges as GPUs, which become more notable for more complex networks with a large number of neuron layers. Furthermore, for digital implementation platforms, the raw input data usually need to be converted to the electrical domain, digitized and processed. Often, a large memory unit is required to store the dataset, which limits the processing time and, in the case of image or video classification, may present privacy implications.

The large bandwidth available at optical frequencies as well as low propagation loss of nanophotonic waveguides (serving as interconnects) make photonic integrated circuits a promising platform to implement fast and energy-efficient processing units[15–17,22] that can augment the performance of conventional digital processors. Recently, photonic implementations of deep neural networks have been reported[15–17,23–30] that offer key features, such as high-speed linear operation and low-loss high-bandwidth connectivity within the network. However, all demonstrations of neural networks so far have been limited to either benchtop setups[29–34] or integration of parts of a deep learning network[15–17,23–28] and, owing to the lack of scalable on-chip non-linear functionality and uncompensated loss of cascaded photonic devices, no scalable, fully integrated photonic deep learning system for data classification has been demonstrated.

Here we report the demonstration of the first integrated end-to-end PDNN that uses computation by propagation to perform sub-nanosecond image classification. Target images are formed on an array of grating couplers serving as input pixels, in which the optical waves impinging on different pixels are coupled into the corresponding nanophotonic waveguides and processed as the light propagates through neurons of different layers on the PDNN chip. Through uniform distribution of a supply light, all neurons in the network have the same optical output range, allowing scalability to a large number of layers. As a proof of concept, the PDNN chip was used for two-class and four-class classification of handwritten letters, achieving accuracies higher than 93.8% and 89.8%, respectively. Measurements show that the PDNN system is capable of achieving an end-to-end classification time of 570 ps, which is comparable with a single clock cycle of

[1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: firooz@seas.upenn.edu

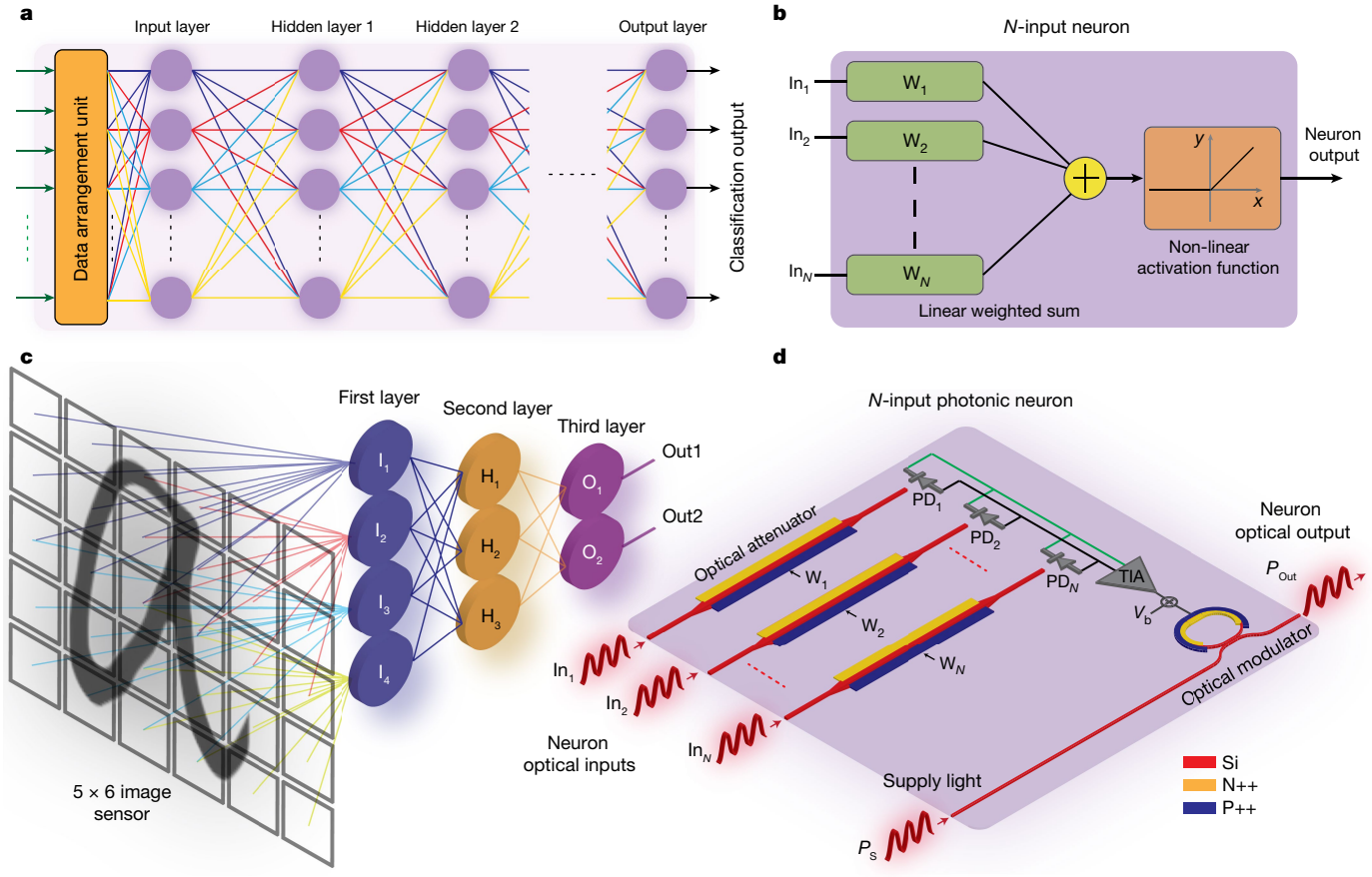


Fig. 1 | Conventional and photonic–electronic deep neural networks. **a**, Block diagram of a conventional deep neural network consisting of a data arrangement unit, followed by the input layer, several hidden layers and an output layer providing classification outputs. **b**, The structure of a conventional *N*-input neuron used in the network in **a**, in which the linear weighted sum of the inputs is passed through a non-linear activation function to generate the neuron output. **c**, The architecture of the implemented PDNN chip, in which the input image is formed on a 5 × 6 pixel array and is arranged into four overlapping sub-images. Pixels of sub-images are routed to the neurons of the first layer. The second and third layers are fully connected to their previous layers. The network generates two outputs. **d**, The structure of an implemented *N*-input photonic neuron, in which the weights of *N* optical input signals are adjusted using optical PIN attenuators and summed after photodetection using parallel PDs. The photocurrent i_{sum} is converted to a voltage and amplified using a TIA. The TIA output is then used to drive an optical MRM realizing the ReLU non-linear activation function, in which the neuron optical output is generated by modulating the supply light.

state-of-the-art digital platforms[35]. As a point of comparison, a conventional deep neural network classifier implemented in the Python environment using Keras[36] achieves 96% accuracy for the same dataset. The implemented PDNN features direct, clock-less processing of input images, which eliminates the need for photodetection, scaling and amplification, analogue-to-digital conversion, data alignment and a large memory module, allowing the realization of much faster and more energy-efficient, yet privacy-aware, neural networks for the next generations of deep learning systems. The PDNN chip was integrated within a footprint of 9.3 mm^2.

The architecture of the implemented PDNN chip and the structure of an *N*-input photonic neuron within the PDNN chip are shown in Fig. 1c, d, respectively. The target image is formed on the input 5 × 6-pixel array, which is divided into four overlapping 3 × 4-pixel sub-images (routings shown in dark blue, red, light blue and light green in Fig. 1c). Input nanophotonic waveguides are arranged to route pixels of each sub-image to a 12-input neuron in the input layer, forming a convolution layer[6,7]. Convolution layers are commonly used within a deep network in image/pattern recognition applications, allowing a lower number of connections and a more efficient feature extraction[8–10]. The outputs of the first layer are fully connected to the three neurons of the second layer. Similarly, the three outputs of the second layer are fully connected to the two neurons of the third layer, generating two network outputs, Out1 and Out2.

The structure of a photonic neuron with *N* optical inputs (In_i) and one optical output is shown in Fig. 2a, in which linear computation (that is, the weighted sum of the input signals) is performed optically and the non-linear activation function is realized opto-electronically. First, an array of 500-µm-long P-doped–intrinsic–N-doped (PIN) current-controlled attenuators is used to individually adjust the optical power in each input nanophotonic waveguide of the neuron. The cross section of the PIN attenuator as well as its microphotograph are shown in Fig. 2b. By forward biasing the PIN junction and injecting carriers, the power of the optical wave (that is, the signal weight) of each neuron input can be adjusted (Fig. 2c). To add the weight-adjusted signals, the outputs of attenuators are photodetected using silicon–germanium (SiGe) photodiodes (PDs) and the resulting photocurrents are combined to generate the weighted sum of the neuron inputs, i_{sum}. The microphotograph of the SiGe PD is shown in Fig. 2d.

To generate the neuron output, the weighted sum of the neuron inputs is passed through a non-linear activation function. Here the rectified linear unit (ReLU) function, offering fast convergence[11,12], is used as the non-linear activation function and is realized by using the electro-optic non-linear response of a PN junction micro-ring modulator (MRM)[37] (Fig. 2e). In Fig. 2a, the electrical current, i_{sum} (that is, the weighted sum of the inputs), is amplified and converted to a voltage using a linear transimpedance amplifier (TIA). The input voltage of

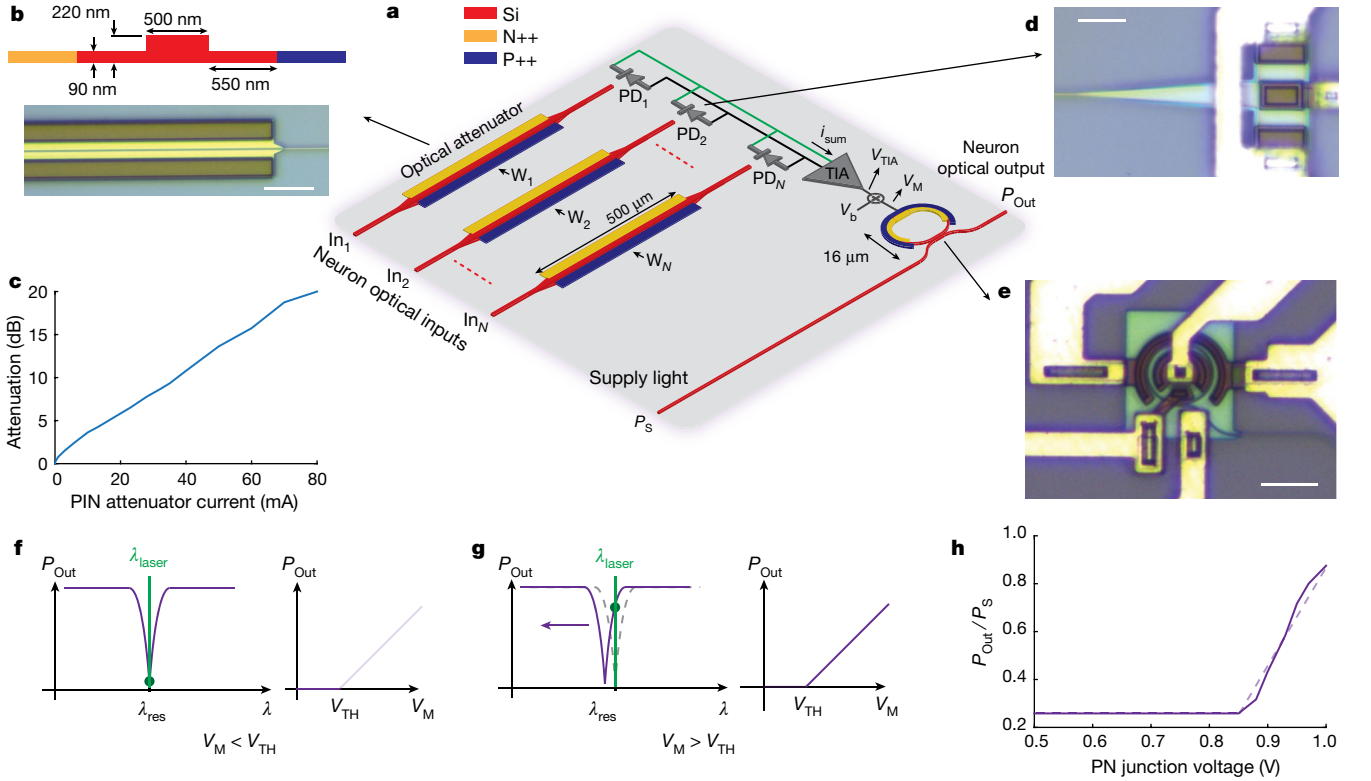


Fig. 2 | Photonic–electronic neuron implementation. a, Schematic of the implemented on-chip photonic–electronic neuron with *N* optical inputs and one optical output, realized using different electro-optical devices. **b**, The cross section and microphotograph of the PIN attenuator, which is realized by creating P++ and N++ doping regions on the two sides of a nanophotonic waveguide. The scale bar is 20 μm. **c**, The attenuation of the PIN attenuator as a function of the injected current. **d**, Microphotograph of a SiGe PD used after each PIN attenuator. The scale bar is 15 μm. **e**, Microphotograph of the MRM used to realize the ReLU activation function. The scale bar is 15 μm. **f**, For the case that the micro-ring is aligned with the wavelength of the supply light, when the voltage across the PN junction (V_M) is smaller than the turn-on voltage of the PN junction (V_{TH}), the junction remains off. In this case, no carriers are injected into the junction and the micro-ring resonance remains unchanged, resulting in a low neuron output power. **g**, For the case that $V_M > V_{TH}$, the PN junction turns on, injecting carriers into the junction. As a result, the waveguide refractive index changes, shifting the micro-ring resonance. In this case, the neuron output power increases as V_M (corresponding to the weighted sum of the neuron inputs) increases. **h**, The measured output power of the MRM (normalized to the supply light power) as a function of the voltage across the micro-ring PN junction, V_M.

the MRM (driving the forward-biased PN junction), V_M, is generated by adding a DC voltage, V_b, to the TIA output voltage, V_{TIA}. The power of a laser, coupled into the chip, is equally distributed among all neurons (within all layers), providing the supply light to the input of the MRM in each neuron. Consider the case that the resonance wavelength of the MRM, λ_{res}, is initially aligned with the wavelength of the supply light, λ_{laser}. When the input voltage to the MRM, V_M, is smaller than the threshold voltage, V_{TH} (corresponding to the built-in potential of the micro-ring PN junction), the PN junction remains off and no carriers are injected into the PN junction (Fig. 2f). Thus, λ_{res} remains aligned with λ_{laser} and the neuron optical output power, P_{Out}, remains low, as the supply light is filtered by the notch response of the MRM. When i_{sum} is large enough such that V_M exceeds V_{TH}, the PN junction turns on and the injected carriers change the refractive index of the optical waveguide in the PN junction. As a result, λ_{res} shifts and the neuron optical output power increases (Fig. 2g). The measured response of the MRM, configured as an electro-optic ReLU, is shown in Fig. 2h, in which P_{Out}/P_S closely follows a rectified linear characteristic as a function of V_M. Note that the ReLU threshold (V_{TH}) can be adjusted by setting V_b.

Photonic deep neural network image classifier chip

Figure 3 shows the top-level architecture of the PDNN, as well as microphotographs of the main blocks. The image of the target object is formed on the 5 × 6 array of input grating couplers serving as the input pixel array. The 30 signals received by the input pixels are split into four sets of overlapping 12-pixel sub-images, each routed to a single neuron of the first layer (I_1 to I_4) using a photonic distribution network, which consists of nanophotonic waveguides, Y-junction splitters and waveguide crossings (Fig. 3b). The details of the image formation are presented in Methods and Extended Data Fig. 1.

A secondary identical on-chip 5 × 6 grating coupler array is used for training of the PDNN chip and the image formation calibration, in which the optical power received by each pixel is monitored using a photodetector (Fig. 3c). The microphotograph of the 5 × 6-pixel array, with an aperture size of about 140 μm by 150 μm, is shown in Fig. 3d.

After the arrangement of the input pixels to overlapping sub-images used to perform convolution, the light is processed using three layers; the first layer (input layer, Fig. 3e), consisting of four 12-input neurons (I_i), is fully connected to three four-input neurons (H_i) of the second (hidden) layer (Fig. 3f). Neurons of the hidden layer are fully connected to the output layer. The third layer (output layer, Fig. 3g) consists of two three-input neurons (O_i). Outputs of I_1 to I_4 and H_1 to H_3 are connected to MRMs R_1 to R_4 and R_5 to R_7 through TIAs/drivers, respectively. The output layer consists of two neurons and, therefore, the classifier allows for two simultaneous outputs (Out1 and Out2) that can be used for up to four-class classification. A laser is coupled into the PDNN chip to provide the supply light to individual neurons in all layers. Figure 3h shows the PDNN chip microphotograph implemented in the AMF 180-nm silicon-on-insulator (SOI) process. The details of chip fabrication and characterization of different devices are included in Methods and Extended Data Table 1.

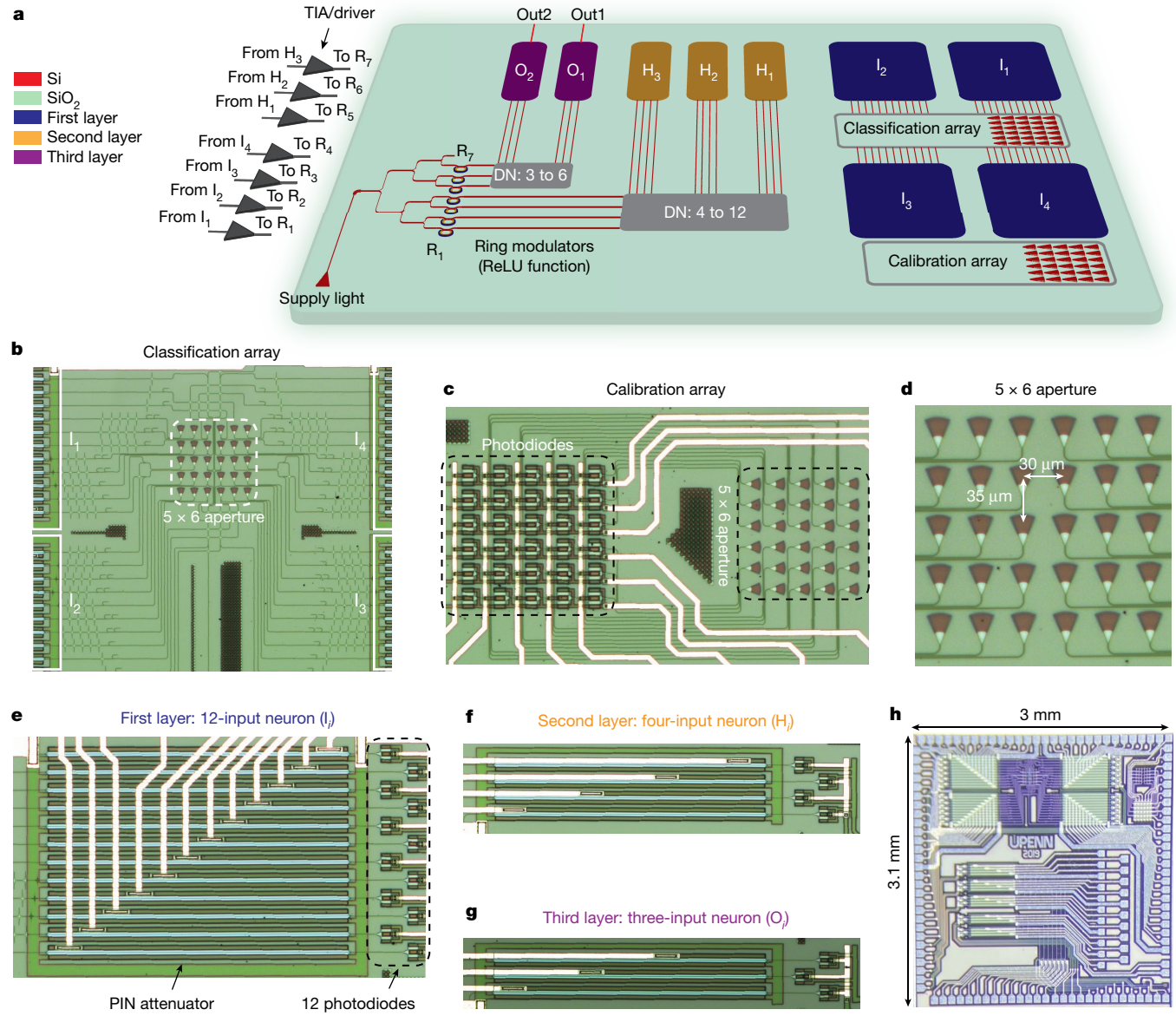


Fig. 3 | The implemented photonic classifier chip. a, The top-level block diagram of the PDNN chip. Two 5 × 6 arrays of grating couplers are used as the input pixel array (**b**) and the calibration array (**c**). **d**, The 5 × 6 array of grating couplers showing the corresponding element pitch. The input pixel array (used for classification) generates four sets of 12 optical signals that are routed to the neurons of the first layer. The supply light is uniformly distributed among the neurons of the second and third layers and passes through seven MRMs to realize the ReLU non-linear activation function. Seven off-chip TIAs are used to drive the on-chip modulators. The system generates two outputs that are used for up to four-class classification. **e**–**g**, Microphotographs of an individual neuron within the first, second and third layers showing the PIN attenuators and the parallel PDs placed after the attenuators, respectively. **h**, Microphotograph of the photonic chip implemented in the AMF 180-nm SOI process. DN, distribution network.

Image classification demonstration

The implemented PDNN chip was used to demonstrate two-class and four-class classifications on two sets of handwritten letters printed on transparency films (Fig. 4a, b). The two-class dataset consists of 216 letters per iteration, 108 'p' and 108 'd'. Similarly, a dataset consisting of 432 letters of 'p', 'd', 'a' and 't', 108 of each (per iteration), was generated to demonstrate four-class classification. Note that the limited print resolution shown in the zoomed-in views of Fig. 4a, b adds an extra level of random variations to the dataset, making the classification more challenging.

First, in the training phase, the network was trained using a set of training images to determine the weight vectors for neurons of all layers. Then, in the classification phase, a different set of test images was classified using the trained PDNN chip. The details of the classification measurement setup are presented in Methods and Extended Data Fig. 1. The network was trained using the on-chip secondary pixel array, in which the training images were formed, the pixel values were recorded and a simulation platform on the basis of Keras[36], an open-source neural network library written in Python, was used to find the weight vectors off-chip. The architecture of this digital neural network is identical to that of the PDNN chip with ReLU as the non-linear activation function. The training and weight optimization were performed using the stochastic gradient descent algorithm[9] (see Methods and Extended Data Fig. 2).

Although the PDNN chip with two outputs can be used in a conventional way to classify a two-class dataset, a simple extra step enables the implemented PDNN chip to perform classification of datasets with a larger number of classes. For example, the difference of the two outputs of the third layer, V_{out} = Out1 − Out2, can be formed and compared with one or a set of threshold values to determine the class of each input

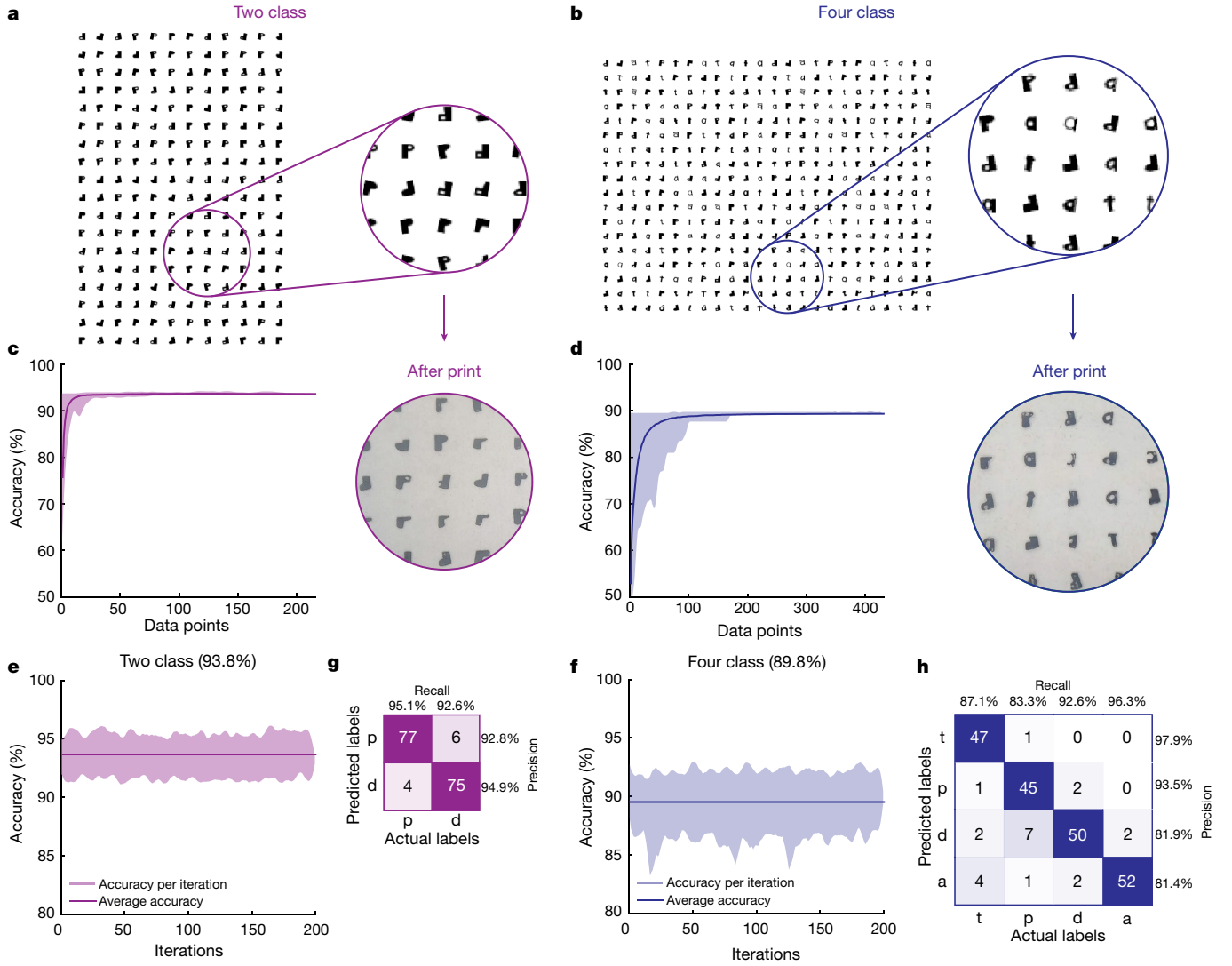


Fig. 4 | Image classification demonstration. a, The dataset consisting of letters 'p' and 'd' for the two-class classification measurements. The zoomed-in views show the letters before and after printing. **b**, The four-class dataset consisting of letters 'p', 'd', 'a' and 't' and the actual image of the printed letters. Classification accuracy during the training process for the two-class (**c**) and the four-class (**d**) cases as a function of input data stream averaged over several iterations during the training process. The shaded area in each graph shows the variation in the accuracy when the input data stream is randomized. Measured classification accuracy as a function of the number of iterations for the two-class case (**e**) and the four-class case (**f**). The confusion matrix, precision and recall corresponding to the two-class case (**g**) and the four-class case (**h**).

image. In this case, the training process also includes the calculation of the threshold values required to optimally separate different classes. In this work, one and three threshold values were used for the two-class and four-class cases, respectively (see Methods and Extended Data Fig. 2).

Figure 4c, d shows the classification accuracy as a function of the number of measured data used to determine the threshold values in the training phase for the two-class and four-class cases, respectively. As shown, the accuracy increases and converges to its maximum value as more data are fed into the algorithm and more precise threshold values are calculated. The calculated threshold values depend on the sequence of the input data. Therefore, to ensure the robustness of the threshold calculation algorithm, this process is repeated several times during the training phase. The shaded areas correspond to the variation in accuracy as a function of the number of input data points. On the basis of these graphs, 25% of the two-class dataset and 50% of the four-class dataset were used to calculate the corresponding threshold values.

The remaining data in both cases were used in the classification phase. The classification accuracies as a function of the number of iterations (that is, using a randomized sequence of input data) for the two-class and four-class cases were calculated using the cross-validation method[38] and shown in Fig. 4e, f. The shaded areas show the variations in the measured accuracies and the solid lines show the average values. Average classification accuracies of 93.8% and 89.8% are measured for the two-class and four-class cases, respectively. In addition to the classification accuracies, the corresponding confusion matrices, precisions and recalls for both cases are shown in Fig. 4g, h, respectively.

These results show that, even with a larger number of classes (that is, the four-class case) and in the presence of printer-induced variations and noise, the PDNN chip still achieves a high classification accuracy. As a point of comparison (in terms of classification accuracy), we used a standard convolutional neural network, implemented in the Python environment using Keras[36], to classify the same printed four-class dataset. This standard convolutional neural network architecture has been previously used for classification of the MNIST[39] handwritten digits dataset to achieve accuracies higher than 99% (ref. [40]) and is tailored to our four-class dataset. This notably larger network (with more than 190 neurons) achieves a classification accuracy of about 96% for the printed four-class ('p', 'd', 'a' and 't') dataset used in this work.

Discussion and conclusion

In general, the classification speed of the proposed PDNN chip is mainly limited by the bandwidths of the MRMs, the SiGe PD and the TIA, since the processing is performed as the waves propagate within the chip. The propagation time of the entire end-to-end PDNN classifier (that is, direct image formation, optical transfer of the input data to the first layer, and several layers of linear and non-linear operations), which corresponds to the end-to-end classification time, is measured to be about 570 ps. The details of the end-to-end propagation time measurements are included in Methods and Extended Data Fig. 3.

Moreover, a sub-60-ps computation speed for linear operations per layer is measured, which corresponds to a linear computation density and energy efficiency of about 3.5 TOPS mm^{-2} (TOPS: tera operations per second) and 345 fJ OP^{-1} per layer, respectively (see Methods and Extended Data Figs. 3, 4 and Extended Data Table 2). Using commercial SOI fabrication processes that offer monolithic integration of electronic and photonic devices[41], an overall bandwidth of tens of gigahertz can be achieved, allowing sub-100-ps total classification time for a similar PDNN architecture (see Methods and Extended Data Figs. 3, 5).

The PDNN architecture can be scaled to a classifier with a larger number of pixels for ultra-fast classification of higher-resolution images and more complex patterns. The availability of low-loss nanophotonic waveguides and splitters within the PDNN architecture markedly reduces the challenge of signal fan-out and distribution compared with all-electronic implementations. The complexity of routing overlapping sub-images to the neurons of the input layer (to perform convolution) can be addressed either by using a fabrication process with several photonic routing layers[42], allowing for more complex photonic routing, and/or through tiling several pixel arrays. More details on the scalability of the implemented PDNN system are included in Methods and Extended Data Fig. 5. Note that the PDNN architecture based on the proposed photonic neuron is not limited to image classification and can be used for other applications. Once the data to be classified are up-converted to the optical domain (for example, through optical modulation), the proposed architecture can be used to perform ultra-fast classification.

In summary, we have demonstrated the first end-to-end PDNN classifier chip that performs sub-nanosecond image classification through computation by propagation of optical waves, eliminating the need for an image sensor, digitization and large memory modules. Low energy consumption and ultra-low computation time offered by our photonic classifier chip can revolutionize applications such as event-driven and salient object detection[43,44], both as a stand-alone classifier or in conjunction with electronic processors, benefiting from the sub-nanosecond classification of the PDNN chip, as well as the reconfigurability and flexibility of electronic processors.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04714-0.

1. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. & Poggio, T. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 411–426 (2007).
2. Wang, D., Su, J. & Yu, H. Feature extraction and analysis of natural language processing for deep learning English language. *IEEE Access* **8**, 46335–46345 (2020).
3. Ribeiro, A. H. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat. Commun.* **11**, 1760 (2020).
4. Lai, L. et al. Computer-aided diagnosis of pectus excavatum using CT images and deep learning methods. *Sci. Rep.* **10**, 20294 (2020).
5. Yuan, B. et al. Unsupervised and supervised learning with neural network for human transcriptome analysis and cancer diagnosis. *Sci. Rep.* **10**, 19106 (2020).
6. Shin, H. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
7. Tajbakhsh, N. et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging* **35**, 1299–1312 (2016).
8. LeCun, Y. & Bengio, Y. in *The Handbook of Brain Theory and Neural Networks* (ed. Arbib, M. A.) 255–258 (MIT Press, 1998).
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
10. Barbastathis, G., Ozcan, A. & Situ, G. On the use of deep learning for computational imaging. *Optica* **6**, 921–943 (2019).
11. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
12. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th International Conference on Machine Learning* (eds Fürnkranz, J. & Joachims, T.) 807–814 (Omnipress, 2010).
13. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).
14. Li, H., Lin, Z., Shen, X., Brandt, J. & Hua, G. A convolutional neural network cascade for face detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5325–5334 (IEEE, 2015).
15. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
16. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photonics* **15**, 102–114 (2021).
17. Bogaerts, W. et al. Programmable photonic circuits. *Nature* **586**, 207–216 (2020).
18. Moons, B. & Verhelst, M. An energy-efficient precision-scalable ConvNet processor in 40-nm CMOS. *IEEE J. Solid-State Circuits* **52**, 903–914 (2017).
19. Lee, J. et al. UNPU: an energy-efficient deep neural network accelerator with fully variable weight bit precision. *IEEE J. Solid-State Circuits* **54**, 173–185 (2019).
20. Hill, P. et al. DeftNN: addressing bottlenecks for DNN execution on GPUs via synapse vector elimination and near-compute data fission. In *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* 786–799 (IEEE, 2017).
21. Nurvitadhi, E. et al. Accelerating binarized neural networks: comparison of FPGA, CPU, GPU, and ASIC. In *2016 International Conference on Field-Programmable Technology (FPT)* 77–84 (IEEE, 2016).
22. Ashtiani, F., Risi, A. & Aflatouni, F. Single-chip nanophotonic near-field imager. *Optica* **6**, 1255–1260 (2019).
23. Cheng, Z., Rios, C., Perince, W. H. P., Wright, C. D. & Bhaskaran, H. On-chip photonic synapse. *Sci. Adv.* **3**, e1700160 (2017).
24. Tait, A. N. et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).
25. Feldmann, J. et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
26. Miscuglio, M. et al. All-optical nonlinear activation function for photonic neural networks. *Opt. Mater. Express* **8**, 3851–3863 (2018).
27. Jha, A., Huang, C. & Prucnal, P. R. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics. *Opt. Lett.* **45**, 4819–4822 (2020).
28. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
29. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
30. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
31. Bueno, J. et al. Reinforcement learning in a large-scale photonic recurrent neural network. *Optica* **5**, 756–760 (2018).
32. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photonics* **15**, 367–373 (2021).
33. Chang, J. et al. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 12324 (2018).
34. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
35. AMD Radeon™ RX 6700 XT Graphics. https://www.amd.com/en/products/graphics/amd-radeon-rx-6700-xt.
36. Chollet, F. et al. Keras. https://keras.io (2015).
37. Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
38. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Series B Stat. Methodol.* **36**, 111–147 (1974).
39. Lecun, Y. et al. The MNIST dataset of handwritten digits. http://yann.lecun.com/exdb/mnist/ (1999).
40. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
41. Rakowski, M. et al. 45nm CMOS — Silicon Photonics Monolithic Technology (45CLO) for next-generation, low power and high speed optical interconnects. In *2020 Optical Fiber Communications Conference and Exhibition (OFC)* (IEEE, 2020).
42. Fahrenkopf, N. M. et al. The AIM photonics MPW: a highly accessible cutting edge technology for rapid prototyping of photonic integrated circuits. *IEEE J. Sel. Top. Quantum Electron.* **25**, 1–6 (2019).
43. Borji, A., Cheng, M., Jiang, H. & Li, J. Salient object detection: a benchmark. *IEEE Trans. Image Process.* **24**, 5706–5722 (2015).
44. Cheng, M., Mitra, N. J., Huang, X., Torr, P. H. S. & Hu, S. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 569–582 (2015).

Methods

Image formation and classification measurement setup

Using the calibration array, we can verify the image formation quality. One important consideration is the uniformity of the image. As shown in Fig. 3b, to ensure uniform optical power distribution, the number of Y junctions and crossings are balanced for all 48 optical paths that route the input pixel array to the first layer of the PDNN chip. To confirm this, we uniformly illuminate the chip, with no obstruction, while measuring the photocurrent of individual pixels in the calibration array. In this case, the measured non-uniformity is less than 5%, which is low enough for the classification system to extract the features properly.

The same measurement can be used to estimate the path loss from the optical collimator output to the output of each pixel (grating coupler). In the case that the power coupled to the collimator is about 63 mW, the measured photocurrent of each PD is about 3 μA. The responsivity of the PD is about 0.8 A W^{-1}. Therefore the power coupled into the waveguide at the output of each grating coupler is estimated to be 4 μW. This results in a total path loss of about 42 dB. This loss is mainly owing to the image formation and coupling loss (not within the network) and includes three factors: (1) the overlap between the input pixel array aperture area and the beam spot, which can be written as

$$\frac{A_{\text{aperture}}}{A_{\text{beam}}} = 0.035 \text{ (equivalent loss} \approx 14.5\text{dB)},$$

(2) pixel-to-aperture ratio (PAR) (that is, the area of each grating coupler relative to the aperture area in Extended Data Fig. 1a), written as $\text{PAR} = \frac{\text{Pixel area}}{\text{Aperture area}} = 0.0048$ (equivalent loss ≈ 23dB),and (3) the grating coupler measured loss of about 5 dB. Note that the transmission coefficient of the transparency film is almost one.

To maximize the efficiency of the input grating couplers (pixels), the carrier board is placed such that the angle between the normal vector of the pixel array plane and the impinging collimated beam is about 12° (Extended Data Fig. 1a). Given that the width of the input grating coupler array is about 150 μm, a 12° angle between the impinging beam and the normal vector of the pixel array plane results in less than 0.1 ps delay between the time that the first grating coupler is illuminated and the time that the last grating coupler is illuminated. Also, the length difference between the shortest waveguide path and the longest waveguide path in Fig. 3b is less than 600 μm, corresponding to a time delay of under 6 ps, which is negligible compared with the per-layer computation time. In future designs with a smaller computation time, this waveguide length mismatch can be greatly reduced by matching the waveguide length in the layout.

Extended Data Figure 1b shows the schematic of the measurement setup used to demonstrate image classification. Two laser sources are used; laser 1 is used for image formation on the classification/calibration arrays and laser 2 is used to provide supply light for neurons. The output power of laser 1, emitting at 1,532 nm, is amplified using an erbium-doped fibre amplifier to approximately 63 mW and coupled to an optical collimator with a beam diameter of 870 μm. The collimated beam illuminates the object plane that consists of printed letters on a transparency film. The printed letters are attached to a custom-fabricated Plexiglas holding frame, which is mounted on a high-precision XY positioning system with a resolution of better than 1 μm. The collimated beam passes through the object plane, forming the image of the target object on the 5 × 6 classification array.

Laser 2 emits 2.5 mW at 1,559.93 nm. Control loops are used to achieve and maintain correct alignment of the MRMs in the presence of thermal and fabrication process variations, resulting in the reliable realization of a rectified linear function. Once all the weights are set and the chip reaches thermal equilibrium, the alignment control loop is engaged to thermally tune the ring modulators, such that all resonance wavelengths are aligned with the wavelength of laser 2. The details of the alignment algorithm are presented in the section 'MRM alignment algorithm' and Extended Data Fig. 6. After all MRMs are aligned, the system sequentially goes through the test images (all printed on the same transparency film) using the XY positioner, while the voltages of Out1 and Out2 ports are continuously monitored. The details of the control circuit are presented in the section 'Electronic control circuitry' and Extended Data Fig. 7. The list of the measurement equipment is provided in Extended Data Table 3.

PDNN chip training process

Before performing the image classification on the test set, the PDNN chip was trained to find the optimal weight vectors. First, the image of each letter in the training set was formed on the secondary on-chip (calibration) pixel array and the corresponding pixel values were recorded. Then, to find the optimal weight vectors, the recorded pixel values for each image of the training set were fed into a digital neural network implemented in Python using Keras[36], with an identical architecture to the PDNN chip and ReLU as the non-linear activation function.

By comparing a simple linear combination of the outputs (that is, V_{out} = Out1 − Out2) to a set of threshold values, we were able to perform two-class and four-class classifications. Therefore, as part of the training process, the threshold values required to optimally separate different classes were calculated. During the training phase, a subset of the data was used for the algorithm to find optimal threshold values. As shown in Extended Data Fig. 2, as the measured data values are fed to the algorithm, the threshold values are constantly revised and move closer to the optimal values, resulting in higher classification accuracies. The optimum weight vectors were translated to the corresponding input voltages of the PIN attenuator array using a lookup table (containing the amount of attenuation as a function of the attenuator input voltage). A microcontroller followed by an array of digital-to-analogue converters (DACs) were used to write the optimum weight vectors into the PDNN chip. During the classification phase, the threshold values calculated during the training process were used.

Classification speed measurements

The classification time of each layer within the PDNN is fundamentally limited by the opto-electronic conversion (using a PD followed by a TIA) and electro-optic conversion (using a ring modulator), that is, $T_{\text{OEO}} \approx 1/BW_{\text{OEO}}$, in which T_{OEO} and BW_{OEO} are the conversion time and bandwidth (set by the PD, TIA and ring modulator), respectively. Hence the total classification time in an N-layer network is limited to $N \times T_{\text{OEO}}$. To demonstrate the capability of the PDNN chip, we conducted a set of measurements as shown in Extended Data Fig. 3 to find the total propagation delay of the chip corresponding to the end-to-end classification time. The propagation delay for each layer is measured as well.

To measure the end-to-end delay, the input light is modulated by a Gaussian monocycle pulse, which is detected after going through the PDNN chip (and off-chip integrated amplifiers). In Extended Data Fig. 3a, the output of laser 1 is modulated by a Gaussian pulse using an intensity modulator. The modulated light is amplified and used to illuminate one of the input pixels. The coupled light is photodetected using a PD of the photonic neuron of the first layer. The PD current is wire-bonded to an amplifier chip (Amp 1), in which it is amplified and converted to a voltage. The amplifier output is wire-bonded back to the PDNN chip. The 3-dB bandwidth of the amplifier is more than 12 GHz. The output of the amplifier drives the forward-biased on-chip MRM of the first-layer neuron to modulate the corresponding supply light. The ring modulator output is routed on-chip to the input of the second layer of the deep network. Similarly, the photocurrent generated in the second-layer neuron is amplified by Amp 2 and drives the MRM in the second layer to generate the input optical signal to the third layer of the PDNN chip. The output photocurrent of the third layer is monitored on an oscilloscope to detect the time of arrival of the Gaussian pulse and, hence, the end-to-end delay of the PDNN classifier. Extended

Data Figure 3a shows a photograph of the packaged PDNN chip and two amplifier chips. Short bond wires are used to minimize the delay of interconnects between the PDNN chip and the amplifier chips.

To de-embed the delay of the off-chip components (intensity modulator, optical amplifier, fibres and RF cables), the exact same setup is used with an on-chip test structure for calibration (Extended Data Fig. 3b). The test structure consists of a grating coupler connected to a single PD through a short (200-μm) waveguide. Note that a fixed direct electrical connection between the pulse generator and the oscilloscope is used as the time measurement reference and, in all of the measurements, the time and amplitude settings of the oscilloscope are kept the same to avoid any measurement errors. Using this setup, we carefully calibrated the effect of all off-chip measurement equipment and devices and, by repeating each measurement several times, ensured the accuracy and repeatability of the measured propagation time. Extended Data Figure 3c shows the measured pulses at the output of the PDNN chip (node A) and the test structure (node B).

The end-to-end processing time (propagation delay) of the PDNN chip (including the two amplifiers) is measured to be about 570 ps. For this measurement, the beginning of each pulse (zero crossing) is used as the point of comparison. Note that the shape of the detected pulse differs from that of the original pulse owing to the non-linear response of the forward-biased MRM. This is confirmed by measuring the response of a single MRM to a Gaussian monocycle pulse, as shown in Extended Data Fig. 3d.

In addition to the end-to-end propagation delay measurement, we also measured the delay of different layers of the PDNN chip without the external amplifiers, using a similar approach as shown in Extended Data Fig. 3. As a result, the total delay of the on-chip photonic circuits that corresponds to the minimum possible classification time is measured to be about 425 ps. The extra 145-ps delay shown in Extended Data Fig. 3c is owing to the bandwidth of the two amplifiers. Also, the propagation delay (computation time) of the linear weight-and-sum block within the first, second and third layers of the PDNN chip are measured to be approximately 46 ps, 58 ps and 56 ps, respectively.

The amplifiers and the photonic devices and blocks of the PDNN can be co-designed and co-integrated on the same chip using fabrication processes that provide both electronic and photonic devices, such as the GlobalFoundries 45CLO process offering monolithic integration of photonics and electronics with PD bandwidth of 50 GHz and transistor f_T (unity current gain bandwidth) of 280 GHz (ref. [41]), allowing BW_{OEO} larger than 40 GHz (T_{OEO} less than 25 ps).

In general, the TIA/amplifier bandwidth can be designed on the basis of the application. Wider band amplifiers can be used for applications that require a lower computation time. Although we have experimentally demonstrated the capability of the implemented PDNN for very fast end-to-end classification (that is, 570-ps classification time), for the two-class and four-class classification demonstrations in this work, because the input image change is performed by physically moving the object plane (which takes about a few hundreds of milliseconds per image), we intentionally did not use wideband TIAs to avoid unnecessary increase in the power consumption. In this case, the low-speed mode, for an end-to-end classification time of under 1 μs, the total power consumption is reduced from 3.75 W (for the high-speed mode at 570-ps end-to-end classification time) to about 2 W.

Speed and energy efficiency comparison with state-of-the-art implementations

Extended Data Figure 4 provides an architecture-level comparison between this work, an all-electrical deep network image classifier and a few other recently demonstrated hybrid optical neural networks. An end-to-end classifier can be implemented using different approaches. As shown in Extended Data Fig. 4a, a conventional end-to-end classifier (in this case, image classifier) typically consists of: (1) an input sensing system (for example, a camera as the input image sensor), (2) a data conversion and compression unit to provide the processor with the proper data type (often through a memory module used to store the input images), (3) a data transfer block and (4) a processor that performs both linear and non-linear operations, which is typically implemented using a GPU and/or an ASIC, such as a tensor processing unit (TPU). However, for state-of-the-art all-electronic deep networks, in which the level of complexity and number of neurons and layers could be very large, typically only the power consumption and time for computation/operations are considered and the power consumption and area of the camera, analogue-to-digital conversion, high-data-rate data transfer (to and from the memory module) and the efficiency of the power supply are not considered nor reported. For example, the Google Edge TPU ASIC (an 8-bit fixed-point system), as a state-of-the-art electronic benchmark, achieves 4 TOPS at 0.5 pJ OP^{-1} (ref. [45]). However, these performance numbers do not address the area and power consumption of the aforementioned blocks and processes and, hence, do not represent the power consumption, area and classification time of the end-to-end classifier. Recently, end-to-end image classifiers have been implemented by incorporating state-of-the-art Google Edge TPUs with cameras (and required interface system), forming so-called smart cameras[45]. One example of such smart cameras is IMAGO Technologies' Edge AI camera that achieves an end-to-end classification time of about 15 ms (on the basis of a maximum frame rate of 65 frames per second[46]) at an estimated total power consumption of 5 W. Another example of an end-to-end image classifier that consists of a camera, GPU, TPU and a neural processing unit (NPU), as well as the required interface system, is the JeVois-Pro smart machine vision camera that can achieve an 8.3-ms end-to-end classification time[47] at a total power consumption of 12 W.

In terms of photonic classifiers, to the best of our knowledge and despite impressive results, none of the recent optical neural networks, such as works in refs. [28,32–34] (Extended Data Fig. 4c–f), have demonstrated an end-to-end photonic-based classifier. The PDNN system reported here (in Extended Data Fig. 4b) is the first photonic-based end-to-end image classifier that includes an on-chip pixel array for image formation, optical routing networks for direct routing of the optical signals between the layers, photonic neurons (with linear optical matrix multiplication and opto-electronic non-linearity) and provides the output classification results.

Equivalent operations per second and total end-to-end classification time. In the PDNN system, in an N-input photonic neuron, the linear weighted sum of the optical inputs is generated and passed through an opto-electronic ReLU non-linear function to generate the neuron optical output. Owing to the analogue (clock-less) nature of the PDNN and the fact that computation is performed as the signal propagates within the network, we estimate the OPS capability of the blocks of the PDNN system (as well as the end-to-end OPS) by finding the equivalent number of operations that a block (or an end-to-end system) performs given the signal propagation time. This signal propagation time (or, equivalently, the computation time) is carefully characterized by measuring the propagation time of a pulse launched through the entire network (to find the end-to-end classification/computation time) or a part of the network (for example, to find the computation time in a layer of the network).

To generate the weighted sum of the inputs in an N-input neuron, N multiplications and N additions are required. Hence each N-input neuron performs $2N$ operations for linear weighted-sum calculations. Assuming that the ith layer consists of $k_i N_i$-input neurons, then the total equivalent number of operations for the linear weighted-sum calculation within the ith layer can be written as $2 \times k_i \times N_i$. In this case, for the linear vector multiplication in the first layer of the implemented PDNN chip with four 12-input neurons (in Extended Data Fig. 4b), given a measured 46 ps computation time (on the basis of pulse travel time

measurements) and the chip area of 0.6 mm^2 (for each multiplication and addition), the linear computation density of about 3.5 TOPS mm^{-2} is achieved. Note that, although the linear computation density is an important performance metric for a sub-block of the end-to-end deep network, it does not represent the overall end-to-end performance of the classifier. For example, a recent impressive work[32] achieved very large linear computation speeds (for a single layer) of 114 and 240 TOPS (for different networks). However, the total end-to-end classification time for the hybrid system in ref. [32], limited by the image sensor, is 6–8 ms, which is seven orders of magnitude larger than the end-to-end classification time of the implemented PDNN, which is 570 ps (in the high-speed mode). An important advantage of the PDNN system is that, because data are directly processed in the optical domain, complementary metal–oxide–semiconductor (CMOS) image sensors, analogue-to-digital conversion, communication (data transfer) and memory modules (to store the input data) are not required.

An equivalent number of operations for end-to-end classification can be approximated for the reported PDNN. Considering that the ReLU non-linear function typically requires three operations (for bias and if-then-else statement[48]), each N-input neuron performs $2N + 3$ operations. Assuming an m-layer network in which the ith layer consists of $k_i N_i$-input neurons, the total equivalent number of operations for the end-to-end classification can be written as

$$\text{Total operations} = \sum_{i=1}^{m} k_i \times (2N_i + 3).$$

This number divided by the total classification time, τ_{prog} (which is inversely proportional to the bandwidth of the opto-electronic and electro-optic conversions set by the bandwidth of the PD, modulator and amplifier), results in the number of OPS for the end-to-end classification, which is approximately equal to

$$\text{OPS}_{\text{end-to-end}} = \frac{1}{\tau_{\text{prog}}} \times \sum_{i=1}^{m} k_i \times (2N_i + 3).$$

The implemented PDNN chip consists of four 12-input neurons in the first layer, three four-input neurons in the second layer and two three-input neurons in the third layer, resulting in a total number of operations per classification of 153 (the last layer does not include non-linearity). Considering a measured end-to-end classification time of 570 ps (Extended Data Fig. 3), the end-to-end PDNN classifier is capable of performing 0.27 TOPS.

Power consumption. For the reported end-to-end PDNN in this manuscript, the power consumption of the linear computation and the power consumption for the end-to-end system can be studied.

The power consumption for linear weighted-sum operations is mainly limited by the power consumption of the electronically controlled optical attenuators, which is about 990 mW for the entire system. In this case, the linear vector multiplication in the first layer of the implemented PDNN chip, with 46 ps computation time (limited by the bandwidth of the PDs), has an average power consumption of 345 fJ OP^{-1}. This is the same for the second and third layers, as the number of operations and the power consumption scale together.

The power consumption of the end-to-end classification is set by the power consumption of the linear weighted-sum operations, the supply light laser (which, considering a 20% wall-plug efficiency, is about 12.5 mW in the low-speed mode and 300 mW in the high-speed mode) and the power consumption of the electronic control circuits and amplifiers, which is about 2.45 W for the high-speed mode (and about 1.05 W for the low-speed mode). In this case, the end-to-end power consumption of the classifier in the high-speed mode is about 3.75 W. Given the total end-to-end classification time of 570 ps, the end-to-end classifier efficiency is about 14 pJ OP^{-1}.

Comparison with the state of the art. To our knowledge, except for the implemented PDNN herein, other reported photonic based classifiers are not end-to-end and the performance of many essential blocks required for an end-to-end classification (for example, camera/sensor array, analogue-to-digital conversion, data storage and transfer and so on) is not included in the reported performance metrics (that is, process/operation time, power consumption and size).

Extended Data Table 2 summarizes the comparison between this work and other photonic-based works, as well as state-of-the-art all-electrical smart cameras implemented as end-to-end classifiers using Google Edge TPUs[45] and/or CPU/GPU/NPUs[45–47].

Note that, although the computation speed of our system is much higher and its size is much smaller than other photonic works in Extended Data Table 2, the classification speed and efficiency of our chip can be further enhanced by using faster PDs. For instance, PDs with bandwidths of about 50 GHz are now available in some commercial processes (for example, GF45CLO[41]) that enable sub-100 ps total computation time for a similar PDNN architecture. Also, using such a process technology node, the electronic blocks can be monolithically integrated with the photonic devices, substantially reducing the area and power consumption, while increasing the operation bandwidth (by eliminating the packaging and interconnect parasitics).

Scalability and classification speed improvement

There are two main factors that make the proposed PDNN architecture scalable.

The first is the use of a supply light. Except for the first layer, whose input signal is received directly from the target object, all neurons within all layers of the network use a supply light that ensures the same level of optical output for all neurons regardless of their location within any layer of the network. This is an important feature of our work that allows this architecture to be easily scaled to a large number of layers, as the variations of the optical power in one layer will not affect the input optical power to the next layer.

Furthermore, the introduction of the per-neuron supply light as well as electrical amplification in this work relaxes the opto-electrical and electro-optical (OE-EO) conversion efficiency requirements. Extended Data Figure 5a shows the proposed general architecture, in which all layers of the PDNN have their dedicated and independent supply light of SL_i. Note that, for larger networks, higher optical power (or several coupling points, for example, a light source for every m network layers) may be required, but the architecture does not limit the number of layers.

The second factor is monolithic implementation on silicon-photonics-enabled CMOS foundry technology processes. An important reason that we chose a foundry-based CMOS-compatible process and used standard PDK cells was to enable scalability to a network with a large number of neurons and layers. For example, CMOS technology has enabled the realization of tens of millions of efficient opto-electronic conversions in a CMOS image sensor. High yield and near-zero incremental cost of transistors has enabled the use of billions of transistors in many different CMOS processes and can certainly be used for amplifier design as well as control electronics in our system when foundry processes with monolithic electronic–photonic integration capabilities are used. These highly efficient compact amplifiers can to some extent compensate for the optical loss, which is also essential for scalability.

As mentioned previously, the classification time is mainly limited by the bandwidth of the OE-EO conversion and, as such, the PD plays an important role in determining the total propagation (classification) time. In our proof-of-concept demonstration, we have used one PD per neuron input. For larger-scale networks that use neurons with a greater number of inputs, there are several ways to maintain a large OE-EO bandwidth.

With the same approach of using one PD per neuron input, an on-chip electrical interconnect network can be designed to enhance the bandwidth. As shown in Extended Data Fig. 5b, one way is to form

a lumped-element transmission line to absorb the capacitance of PDs to markedly increase the bandwidth (that is, similar to the bandwidth enhancement techniques used in a distributed amplifier[49]).

Although in the PDNN chip the weighted optical inputs are converted to the electrical domain and added to generate the weighted-sum signal (that is, each input of a neuron is photodetected after weight adjustment and photocurrents are combined), they can also be added in the optical domain. In this case, only one PD is required for each neuron. As shown in Extended Data Fig. 5c, after weight adjustment using PIN attenuators, the optical signals are added together and the resulting signal is coupled to one PD. In this way, regardless of the number of neuron inputs, the opto-electronic conversion time is limited by the capacitance of a single PD. Note that such a scheme may require slow optical phase adjustment (using slow phase shifters in Extended Data Fig. 5c) in each input (during a calibration phase) to avoid undesired destructive interference owing to relative optical phase mismatches or a thermal gradient (the latter being less probable, as neuron input waveguides are closely placed in an integrated structure). Such a neuron structure also enables complex signal analysis, with applications such as 3D object classification.

Finally, the PD that is used in the AMF 180-nm process has a parasitic capacitance of about 17 fF (ref. [50]). There are processes available that offer PDs with much smaller capacitance. For instance, the GF45CLO process (a 45-nm technology node) offers PDs with a smaller parasitic capacitance that enable much larger opto-electronic conversion bandwidths, as well as efficient RF/mm-wave amplifiers with a high gain–bandwidth product. In addition, such processes enable monolithic integration of photonic and electronic circuits, which reduces the packaging parasitics.

In summary, the proposed PDNN architecture can be scaled to networks with a large number of layers, a large number of neurons per layer and neurons with a large number of inputs.

MRM alignment algorithm

In the implemented PDNN chip, seven MRMs are used to implement and approximate the neural ReLU non-linear activation function; four MRMs at the output of the first layer and three MRMs at the output of the second layer. As discussed earlier, a supply light is coupled into the optical input of each MRM. Because the wavelength of the supply light is the same for all micro-rings, the resonance wavelengths of all MRMs must be aligned to ensure reliable and repeatable realization of ReLU functions for all neurons. In practice, the resonance wavelength of MRMs may vary owing to fabrication process variations and temperature change. Therefore, in addition to a careful design and layout of the micro-rings, control loops were implemented to compensate for any misalignments between the resonance wavelength of MRMs and the wavelength of the supply light. Each MRM can be tuned using an N-doped heater section (serving as a thermal phase shifter) with a measured resistance of about 1.9 kΩ. Extended Data Figure 6a shows the algorithm used to perform the micro-ring alignment. First, the supply light is switched on and all weights are set; properly setting the weights is of particular importance, as biasing the PIN attenuators may increase the temperature of the chip. Therefore, the micro-ring alignment process should be performed when the weights are set and the chip has reached thermal equilibrium. In this case, the control loop sequentially adjusts the heater voltages to minimize the difference between the sum of the outputs of the neurons of the second and third layers (that is, H_1 to H_3, O_1 and O_2), V_{SUM}, and a reference voltage, V_{REF}. In addition, the algorithm ensures that the heater voltages do not exceed the maximum allowable value, V_{max}. At the end of each iteration (that is, after the adjusting the voltage of all heaters), if V_{SUM} becomes smaller than V_{REF}, then V_{REF} is set to V_{SUM} and the next iteration starts. Once all rings are aligned, the optimal heater-biasing voltages are recorded and used during the classification process.

To verify the performance of the ring alignment control loop, first, the laser that illuminates the 5 × 6-input pixel array is turned off (Extended Data Fig. 6b). In this case, the outputs of the neurons of the first layer (I_i) are zero. Because micro-rings are properly aligned, the outputs of the neurons of the second and third layers (the corresponding ReLU function output) remain low. Then, the input laser is turned on, uniformly illuminating the 5 × 6-input pixel array (Extended Data Fig. 6c). In this case, I_1 to I_4 increase, shifting the resonance wavelengths of the MRMs, which results in a large change in the outputs of the neurons of the second layer, H_1 to H_3. Similarly, the output of the neurons of the third layer, O_1 and O_2, will change. The output voltages of neurons of different layers before and after uniform illumination of the input pixel array are also shown in Extended Data Fig. 6b, c.

Electronic control circuitry

Extended Data Figure 7 shows the block diagram of the electronic system used to control and drive the photonic components of the classifier chip. The circuit consists of a microcontroller used to generate the data and clock signals for the serial DAC array to set the weights, thermally align the MRMs and adjust the threshold voltage of each ReLU block. A serial interface is used to write the data into the serial DAC array. There are 66 PIN attenuators on chip to set the weights corresponding to the neurons (4 × 12 in the first layer, 3 × 4 in the second layer and 2 × 3 in the third layer), seven heaters are used for thermal tuning of the micro-rings and seven bias voltages (V_b) to adjust the threshold of the ReLU blocks.

Chip fabrication

The photonic chip was fabricated in the AMF 180-nm SOI process with a 2-µm-thick buried oxide. Single-mode, 220-nm-thick and 500-nm-wide nanophotonic waveguides with a loss of less than 2 dB cm^{-1} were used for photonic routing. The details of the device performance metrics are provided in Extended Data Table 1.

Data availability

The data that support the plots in Fig. 4 can be accessed at https://doi.org/10.5061/dryad.q2bvq83mw.

Code availability

Codes that are used in this paper are available from the corresponding author on reasonable request.

45. Kist, A. M. Deep learning on edge TPUs. Preprint at https://arxiv.org/abs/2108.13732 (2021).
46. IMAGO Technologies' Edge AI camera. https://imago-technologies.com/wp-content/uploads/2021/01/Specification-VisionAI-V1.2.pdf.
47. JeVois smart machine vision. https://www.jevoisinc.com/collections/jevois-hardware/products/jevois-pro-deep-learning-smart-camera.
48. Kulyukin, V. et al. On image classification in video analysis of omnidirectional *Apis mellifera* traffic: random reinforced forests vs. shallow convolutional networks. *Appl. Sci.* **11**, 8141 (2021).
49. Chiu, T. Y., Wang, Y. & Wang, H. A 3.7–43.7-GHz low-power consumption variable gain distributed amplifier in 90-nm CMOS. *IEEE Microw. Wirel. Compon. Lett.* **31**, 169–172 (2021).
50. Xuan, Z. et al. A low-power 40 Gb/s optical receiver in silicon. In *2015 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)* 315–318 (IEEE, 2015).

Acknowledgements This work was supported by the Office of Naval Research of the United States under award number N00014-19-1-2248.

Author contributions F. Ashtiani and F. Aflatouni conceived the design idea. F. Ashtiani designed, simulated and laid out the photonic chip. F. Ashtiani and A.J.G. conducted the measurements. F. Aflatouni supervised the project. F. Ashtiani and F. Aflatouni wrote the manuscript.

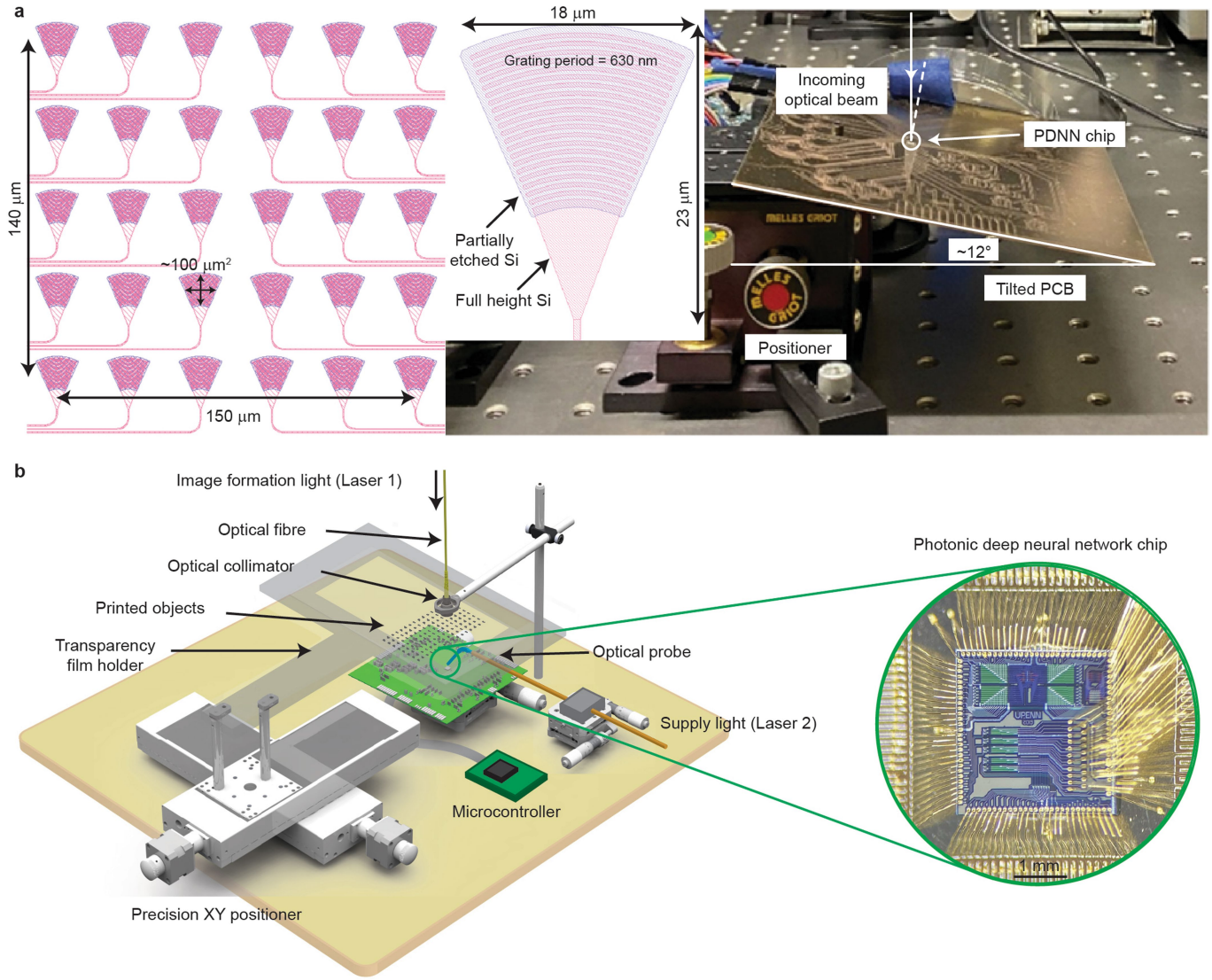
Competing interests F. Aflatouni and F. Ashtiani have filed a patent on the proposed PDNN architecture (publication number WO2022020437A1).

Additional information

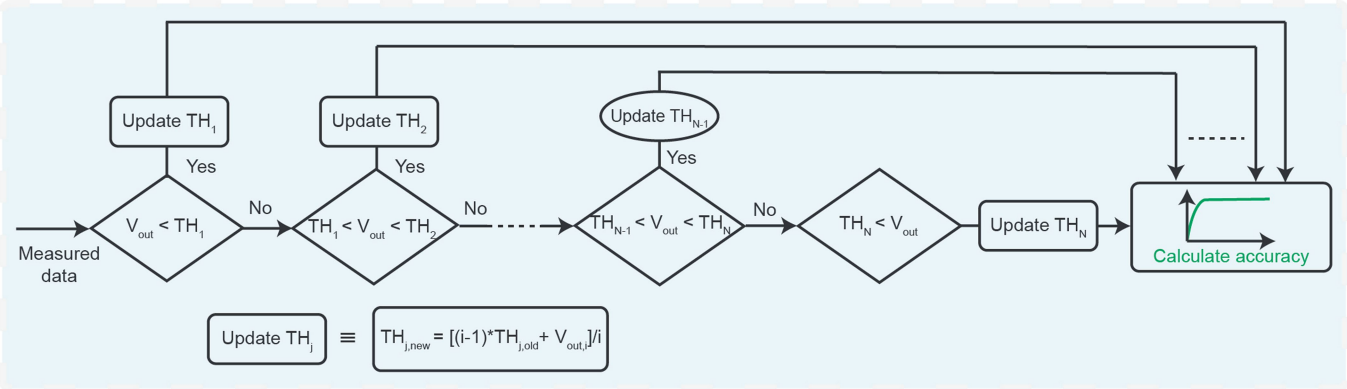
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04714-0.

Correspondence and requests for materials should be addressed to Firooz Aflatouni.

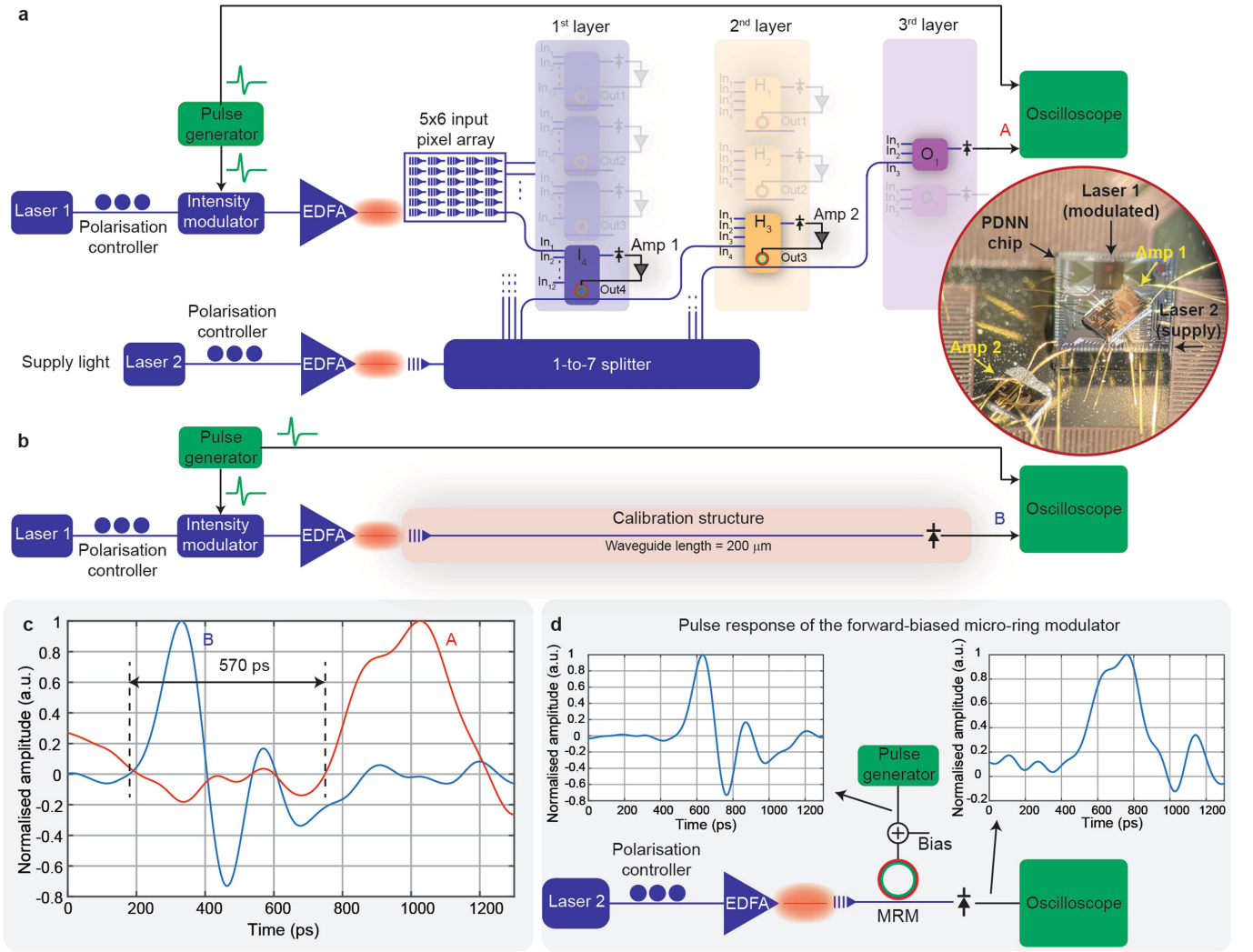
Peer review information *Nature* thanks Wolfram Pernice and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



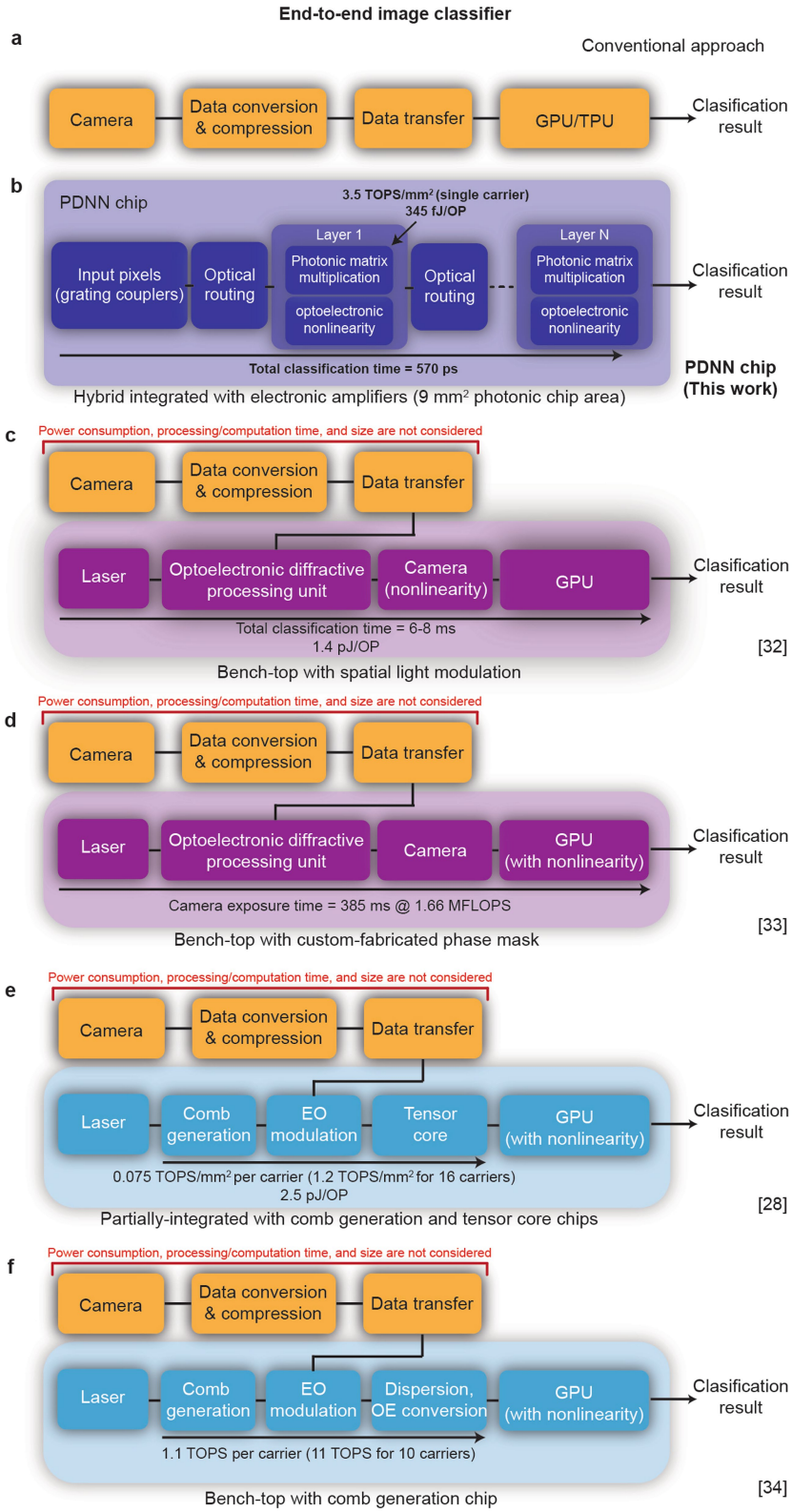
Extended Data Fig. 1 | Image formation and measurement setup. a, The input pixel array aperture and grating coupler (pixel) design are shown. The printed circuit board (PCB) is tilted (by about 12°) to maximize the pixel efficiency at 1,532 nm. **b**, Classification measurement setup. Laser 1, emitting at 1,532 nm, serves as the light source for image formation on the input pixel array (in the classification phase) or the calibration array (in the training phase), whereas laser 2, emitting at 1,559.93 nm, is used as the supply light. The target objects (dataset) are printed on a transparency film mounted on a custom-fabricated frame. A high-precision XY positioner is used for scanning through the dataset. A microcontroller is used to write the weights into the photonic chip and to implement MRM alignment control loops.



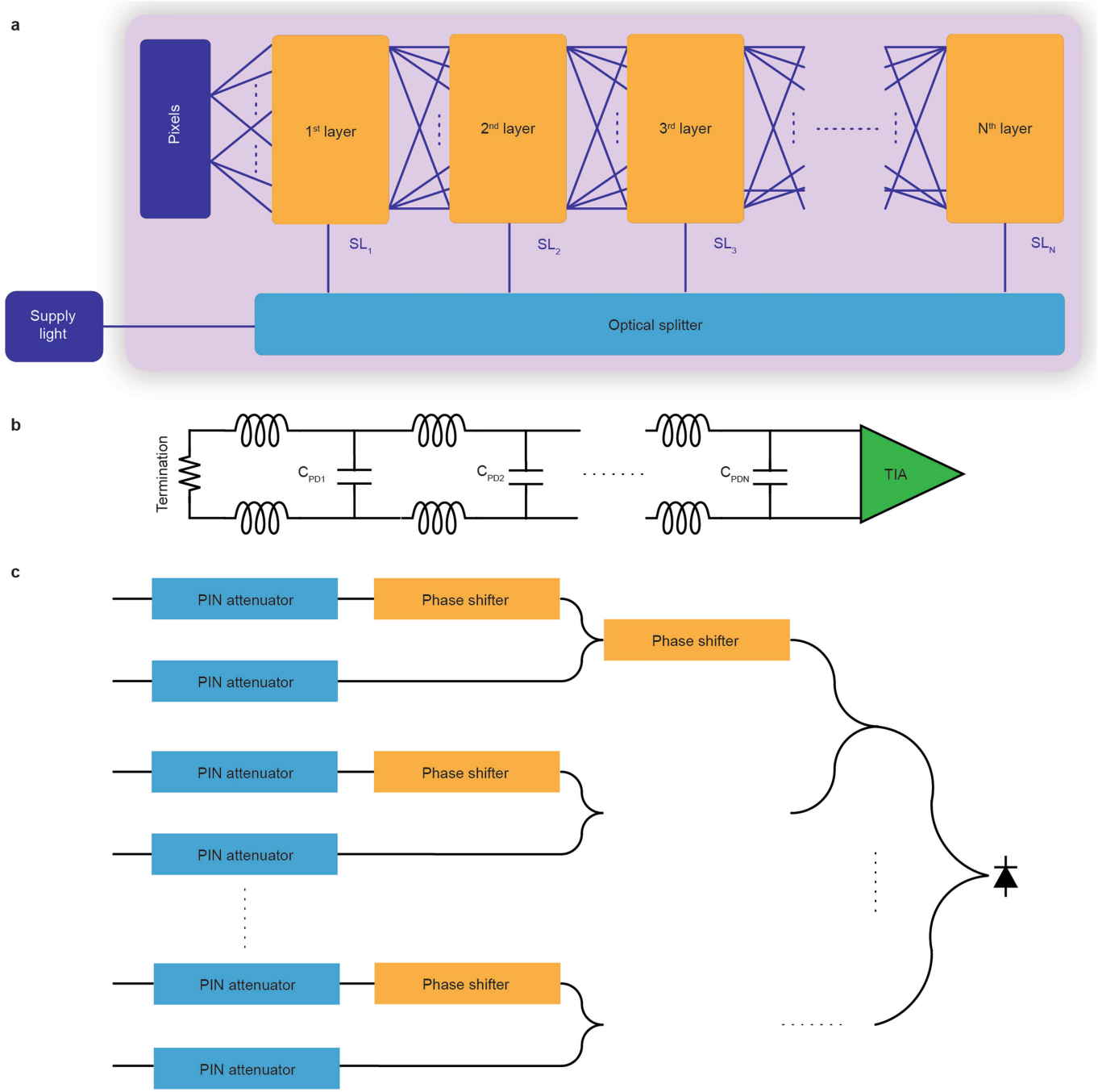
Extended Data Fig. 2 | PDNN chip training and threshold calculations. **a**, The implemented algorithm to find and revise the threshold values to properly separate *N* different classes. A linear combination of the network output, in this case, the differential output defined as V_{out} = Out1 − Out2, is measured and compared with different threshold levels. The threshold values (TH_j) are revised one by one as measured network differential output values ($V_{out,i}$) are sequentially passed into the algorithm.



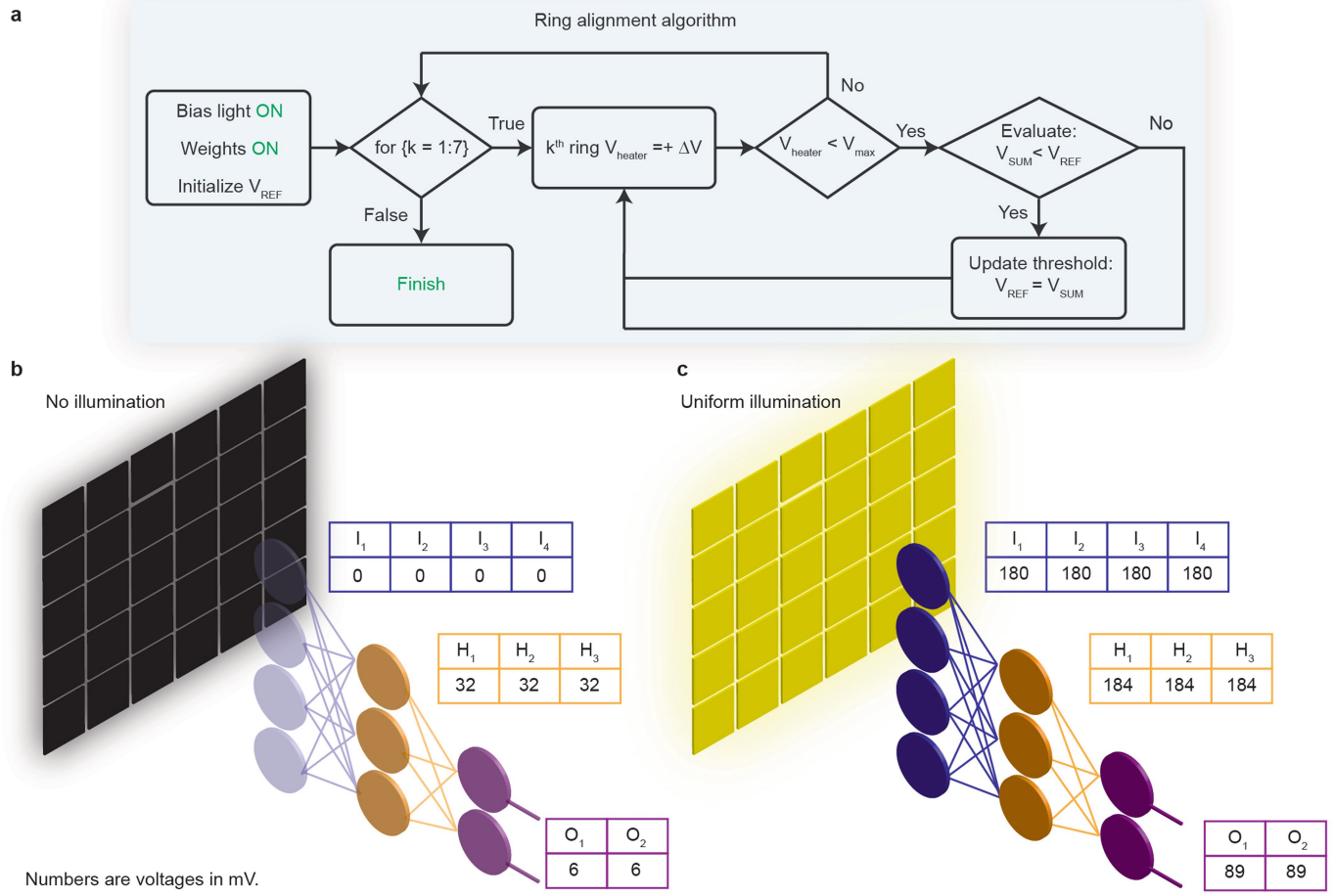
Extended Data Fig. 3 | Propagation time measurement. a, Propagation time measurement setup and packaging. **b**, Calibration setup using a test structure, which consists of a grating coupler and a PD. **c**, Two detected pulses at nodes A and B showing an end-to-end system delay of about 570 ps. **d**, Measurement setup used to show the effect of the forward-biased MRM response on the pulse shape.



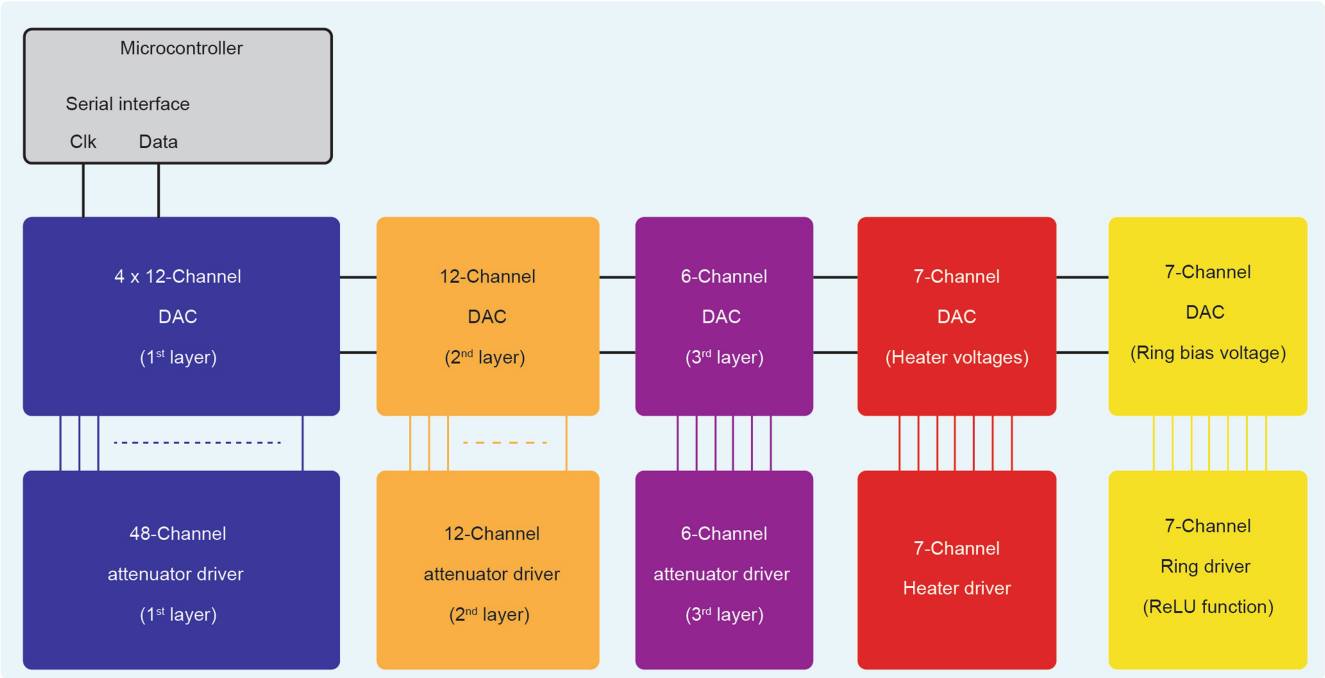
Extended Data Fig. 4 | Comparison with the state of the art. Image classification schemes implemented using optical and electronic neural networks.



Extended Data Fig. 5 | Scalability and computation time enhancement methods. a, An *N*-layer photonic neural network, in which each layer has its dedicated supply light, allowing scalability to a deep network with a large number of layers. Bandwidth enhancement by absorbing the parasitic capacitance of PDs in a lumped-element transmission line (**b**) and using one PD per neuron (after optical combining) (**c**).

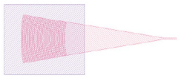
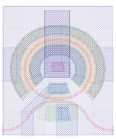
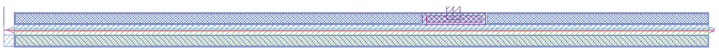
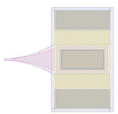


Extended Data Fig. 6 | Micro-ring alignment algorithm and characterization. a, The implemented algorithm flow chart for micro-ring alignment. The cost function to be minimized is V_{SUM}, which is the sum of the outputs of the second and third layers (that is, H_i and O_i). All micro-rings are thermally tuned to find the optimal heater voltages that correspond to the same resonance wavelength for all seven rings. **b**, In case of no input illumination, the outputs of the neurons of the first layer (I_i) are zero. If micro-rings are properly aligned, the outputs of the neurons of the second and third layers remain low. **c**, In the case that the optical input is uniformly illuminating the input pixel array, if all rings are aligned, I_1 to I_4 will increase, shifting the resonance wavelengths of the MRMs, which results in a large change in the outputs of the neurons of the second and third layers.



Extended Data Fig. 7 | Electronic control circuit block diagram. The microcontroller sends the clock and data signals to the serial DACs, whereas the outputs of the DACs are connected to their corresponding drivers to drive the on-chip photonic devices (PIN attenuators, ring PN junctions and micro-ring thermal phase shifters).

Extended Data Table 1 | Performance metrics of the different on-chip devices

Device	Layout	Performance metrics
Grating coupler (pixel)		Coupling efficiency: 30%
Grating coupler (supply light)		Coupling efficiency: 40%
Waveguide crossing		Insertion loss < 0.05 dB Isolation > 37 dB[49]
Y-junction		Excess loss ~ 0.5 dB
Single-mode Si waveguide		Loss < 2 dB/cm
Ring modulator (for ReLU generation)		Input capacitance ~ 30 fF Bandwidth > 30 GHz
PIN attenuator		Insertion loss ~ 0.3 dB
SiGe photodiode		Responsivity: 0.8 A/W Bandwidth ~ 30 GHz

Extended Data Table 2 | Performance comparison with state-of-the-art optical and electronic implementations

Metric	[45,46]	[47]	[32]	[33]	[28]	[34]	This work
Architecture	All electronic (Vision AI smart camera)	All electronic (JeVois-Pro smart camera)	Single-layer optical + GPU	Single-layer optical + GPU	Single-layer photonic + digital	Single-layer photonic + digital	Photonic-mmWave (end-to-end)
End-to-end	Yes	Yes	No	No	No	No	No
Implementation platform	CMOS chip on PCB	CMOS chip on PCB	Bench-top	Bench-top	Photonic chip (single-layer) + digital	Bench-top	Integrated (Packaged with off-chip integrated amplifiers)
End-to-end classification time	15 ms	8.3 ms	6-8 ms*	385 ms*	N/A	N/A	570 ps†
Linear weight-sum operations (TOPS)	4‡	13§	110/240 (single layer)	0.0017	4‖ (0.25 per carrier)	11 (1.1 per carrier)	2.07
Linear operation energy consumption (pJ/OP)	0.5	0.56	1.4	N/A	2.5	N/A	0.345 (per layer)
End-to-end operations (TOPS)	N/A	N/A	N/A	N/A	N/A	N/A	0.27
End-to-end power consumption	5 W	12 W	N/A	N/A	N/A	N/A	3.75 W (14 pJ/OP)¶

* Does not include the camera (image formation), data conversion and data transfer.
† For the end-to-end classifier (in the high-speed mode), including image formation and data transfer. End-to-end classification time in the low-speed mode is under 1μs.
‡ Reported for the Google Edge TPU (not the entire system)[45].
§ Calculated on the basis of reported performance for the Google Edge TPU+GPU+NPU (ref. [47]).
‖ Calculated on the basis of the size of each MAC in ref. [28].
¶ High-speed mode.

Extended Data Table 3 | List of equipment and devices

Equipment	Model
Laser 1	HP 8168F
Laser 2	Agilent 81682B
XY positioner and controller	Thorlabs NRT150 and BSC103
RF amplifier (high-speed)	Analog Devices HMC8412CHIPS
Driver op-amp	Texas Instruments TLV3544
Digital-to-analogue converter (DAC)	Analog Devices AD8802
Microcontroller	ATMEL ATSAM3X8E
Optical collimator	Thorlabs CFC-5C
Polarisation controller	Thorlabs FPC-31
Erbium-doped fibre amplifier	Optilab EDFA-I24-B
DC power supply	Agilent E3646A