

Higher-dimensional processing using a photonic tensor core with continuous-time data

Received: 20 January 2023

Accepted: 17 September 2023

Published online: 19 October 2023

 Check for updates

Bowei Dong^{1,2}, Samarth Aggarwal¹, Wen Zhou¹, Utku Emre Ali¹, Nikolaos Farmakidis¹, June Sang Lee¹, Yuhan He¹, Xuan Li¹, Dim-Lee Kwong², C. D. Wright³, Wolfram H. P. Pernice^{4,5} & H. Bhaskaran¹✉

New developments in hardware-based ‘accelerators’ range from electronic tensor cores and memristor-based arrays to photonic implementations. The goal of these approaches is to handle the exponentially growing computational load of machine learning, which currently requires the doubling of hardware capability approximately every 3.5 months. One solution is increasing the data dimensionality that is processable by such hardware. Although two-dimensional data processing by multiplexing space and wavelength has been previously reported, the use of three-dimensional processing has not yet been implemented in hardware. In this paper, we introduce the radio-frequency modulation of photonic signals to increase parallelization, adding an additional dimension to the data alongside spatially distributed non-volatile memories and wavelength multiplexing. We leverage higher-dimensional processing to configure such a system to an architecture compatible with edge computing frameworks. Our system achieves a parallelism of 100, two orders higher than implementations using only the spatial and wavelength degrees of freedom. We demonstrate this by performing a synchronous convolution of 100 clinical electrocardiogram signals from patients with cardiovascular diseases, and constructing a convolutional neural network capable of identifying patients at sudden death risk with 93.5% accuracy.

Owing to the proliferation of the Internet of Things and 5G, the global data volume has grown exponentially, reaching 64.2 zettabytes in 2020 and is projected to reach 181.0 zettabytes in 2025 (ref. 1). Big data provides machine-learning (ML) models with unprecedented rich and multifaceted information to reveal underlying data patterns for analysis and prediction², with profound societal impact in diverse fields³ such as computer vision⁴, speech recognition⁵, natural language processing⁶, physical sciences⁷, computer sciences⁸ and biomedical

sciences⁹. However, the heavy computational load that big data imposes on hardware systems threatens the viability of ML¹⁰. Matrix–vector multiplication (MVM) is the fundamental operation that dominates 90% of runtime in most ML models (for example, GoogleNet, VGG, OverFeat and AlexNet)¹¹. To parallelize MVM by increasing the dimensionality of data, various electronic computing architectures with the parallel-mode advantage compared with central processing units (CPUs) have been employed in hardware¹², such as graphics processing

¹Department of Materials, University of Oxford, Oxford, UK. ²Institute of Microelectronics, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. ³Department of Engineering, University of Exeter, Exeter, UK. ⁴Institute of Physics, University of Muenster, Muenster, Germany. ⁵Kirchhoff-Institute for Physics, Heidelberg University, Heidelberg, Germany. ✉e-mail: harish.bhaskaran@materials.ox.ac.uk

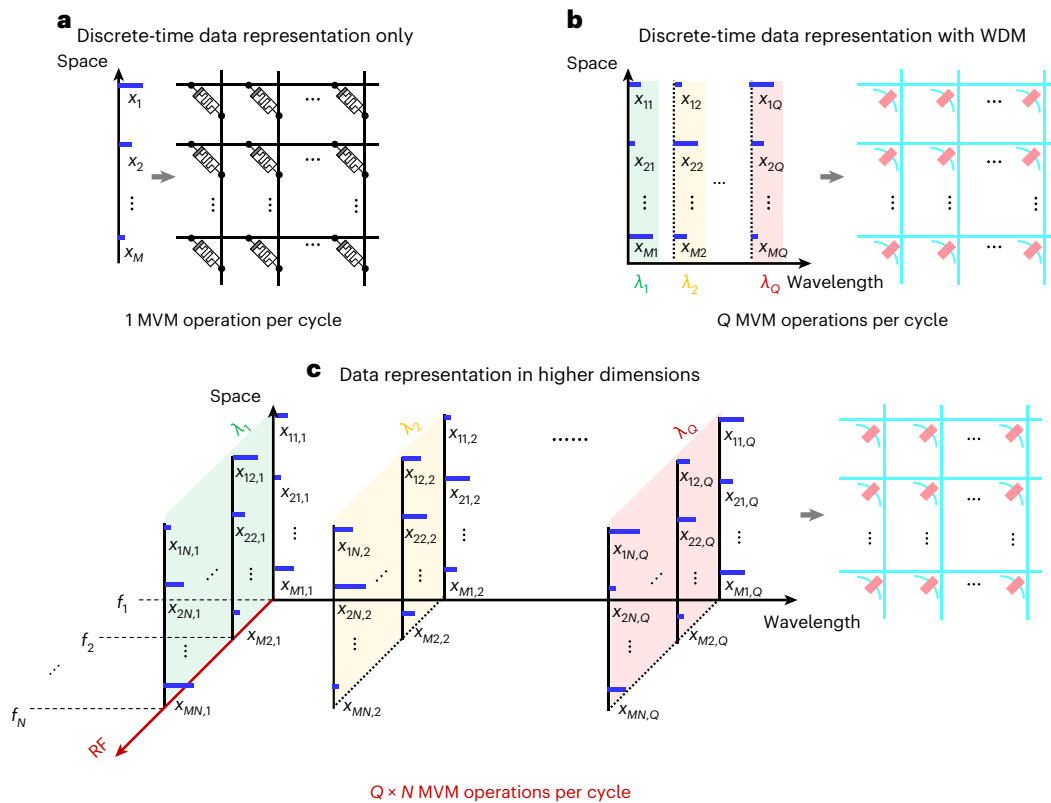


Fig. 1 | High-dimensional photonic in-memory computing using data with three DOFs. Comparison of computing schemes. **a**, Traditional electronic computing uses the spatial DOF for data input, inputting 1D arrays to achieve MVM. **b**, Recent photonic computing uses the spatial and wavelength DOFs,

inputting 2D arrays to achieve matrix–matrix multiplications. **c**, Our scheme adds the RF DOF by using continuous-time data representation, inputting 3D arrays to achieve parallel matrix–matrix multiplications.

units¹³, field-programmable gate arrays¹⁴ and application-specific integrated circuits¹⁵. In addition, perhaps the most notable recent advance is the use of memristive crossbar arrays for analogue in-memory computing^{16–18}. Various mechanisms have been explored to store memories in physical states of materials (redox¹⁹, phase change²⁰, ferroelectric²¹ and magnetoresistive²²) to enable such in-memory computing. A memristive crossbar array with M inputs and K outputs mathematically represents a matrix of dimension $d_{K \times M}$ that contains K $d_{1 \times M}$ kernels. Multiplication and addition operations are performed according to Ohm’s law and Kirchhoff’s law, respectively. The input data use the spatial degree of freedom (DOF) and are a one-dimensional (1D) array $X_{1D} = (x_1, x_2, \dots, x_M)^T$ representing a $d_{M \times 1}$ vector, leading to one $d_{K \times M} \times d_{M \times 1}$ MVM per operation cycle (Fig. 1a).

Photonic MVM is emerging as a next-generation alternative with the advantages of low latency, low energy consumption and high DOFs^{23,24}. Compared with electronic data transmission that is inherently limited by capacitive delay and the energy consumption required to charge/discharge electronic integrated circuits, photons transmit data at the speed of light with near-zero power consumption²⁵. Photonic MVM can access a huge terahertz bandwidth compared with a gigahertz bandwidth accessible by electronics, opening the possibility of high parallelism by exploiting the wavelength DOF, that is, wavelength-division multiplexing (WDM). Traditionally, photonic MVM was implemented by light diffraction in free space, an approach that continues to inspire computing architectures²⁶. In the past decade, photonic MVM using photonic integrated circuits (PICs) has flourished^{27,28} owing to the development of scalable on-chip dense integration of optical waveguide components^{29,30}. Notable progress includes the demonstration of PIC-based MVM processors based on cascaded Mach–Zehnder interferometer arrays using coherent light as the data carriers and thermo-optic phase shifters as weighting

elements³¹. Broadcast-and-weight PIC-based MVM processors using light at different wavelengths as data carriers and tunable microring resonator add–drop filters as weighting elements have also been developed³². More recently, optical frequency comb technology was introduced with PIC-based MVM processors to provide a high-quality multiwavelength light source with dense wavelength spacing^{33,34}. A record high of 11 tera operations per second has been realized using a single optical frequency comb with the wavelength-and-time interleaving technique³³. The latest advance in delocalized photonic deep learning shows the advantages of using PIC-based MVM processors on the Internet’s edge³⁵. In addition, it is worth noting that a photonic counterpart of an electronic crossbar array has been demonstrated³⁴. The passive photonic crossbar array uses waveguide directional couplers and crossings as interconnects and phase-change materials (PCMs) as memories (optical transmissions tuned by the non-volatile crystalline state of the PCM³⁶).

In all the PIC-based MVM processors, two DOFs are accessible by the input data, that is, space and wavelength, allowing a two-dimensional (2D) array input

$$X_{2D} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1Q} \\ x_{21} & x_{22} & \dots & x_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \dots & x_{MQ} \end{bmatrix}$$

(Fig. 1b). Here Q $d_{M \times 1}$ input vectors, each carried by a different wavelength λ_q , can be processed in parallel, leading to one $d_{K \times M} \times d_{M \times Q}$ matrix–matrix multiplication (equivalent to Q $d_{K \times M} \times d_{M \times 1}$ MVMs). A parallelism (defined as the number of MVMs per operation cycle of

a physical device) of 4 using a photonic crossbar array and WDM has been realized³⁴. Recently, a similar endeavour to increase data dimensionality was reported in electronic crossbar arrays by exploring the continuous-time data representation³⁷. Conceptually similar to WDM, continuous-time data are generated by multiplexing radio-frequency (RF) signals at different frequencies, where data are encoded in RF amplitudes. As this was done in electronics, the input data are a 2D array restricted to spatial and RF DOFs, leading to one $d_{K \times M} \times d_{M \times N}$ matrix–matrix multiplication (equivalent to $N d_{K \times M} \times d_{M \times 1}$ MVMs) if N RF components are used. Inspired by such advances, in this paper, we demonstrate a computing architecture in hardware that allows three-dimensional (3D) array inputs for higher-dimensional MVM by simultaneously exploiting three DOFs, that is, space, wavelength and RF. The input data are a 3D array:

$$X_{3D} = [X_{2D,\lambda_1} \ X_{2D,\lambda_2} \ \dots \ X_{2D,\lambda_Q}], X_{2D,\lambda_q} = \begin{bmatrix} x_{11,q} & x_{12,q} & \dots & x_{1N,q} \\ x_{21,q} & x_{22,q} & \dots & x_{2N,q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1,q} & x_{M2,q} & \dots & x_{MN,q} \end{bmatrix}$$

X_{3D} represents multiple $d_{M \times N}$ matrices each carried by a wavelength λ_q , when N RF components (f_1 to f_N) and Q wavelengths are used (Fig. 1c). The 3D array input is processed by an electro-optically controlled photonic tensor core with reconfigurable non-volatile PCM memories to enable photonic in-memory computing. Our system is effectively implementing $Q d_{K \times M} \times d_{M \times N}$ matrix–matrix multiplications (equivalent to $Q \times N d_{K \times M} \times d_{M \times 1}$ MVMs) and achieves a remarkable ultra-high parallelism of 100, two orders higher than the previous implementation³⁴ using only two DOFs. Having such a higher-dimensional processing advantage allows our system to accelerate hugely common artificial-intelligence-type processing tasks. We demonstrate this by realizing the synchronous convolution of 100 clinical electrocardiogram (ECG) signals from cardiovascular disease (CVD) patients and facilitating a convolutional neural network (CNN) to identify patients at sudden death risk with 93.5% accuracy. Increasing the dimensionality from 1D to 2D to 3D data processing by exploiting additional DOFs, the system parallelism is increased from 1 to (Q or N) to $Q \times N$, providing a viable path for ultraparallel photonic computing.

Data architecture and working principle

The proposed computing architecture utilizes continuous-time data representation instead of traditional discrete-time data representation to add RF as the third DOF for data input. Figure 2 conceptually illustrates the data architecture and working principle of using continuous-time data representation. An example of matrix–matrix multiplication without using WDM is illustrated to highlight RF parallelism and maintaining visual clarity (Fig. 2a).

To perform higher-dimensional in-memory computing that simultaneously utilizes the spatial, wavelength and RF DOFs, a photonic tensor core system based on electro-optically controlled PIC technology is proposed (Fig. 2b). To implement the matrix–matrix multiplication shown in Fig. 2a, the photonic tensor core with M inputs and K outputs defines a $d_{K \times M}$ matrix

$$W = \begin{bmatrix} w_{11}w_{21} \dots w_{K1} \\ w_{12}w_{22} \dots w_{K2} \\ \vdots \\ w_{1M}w_{2M} \dots w_{KM} \end{bmatrix}^T$$

consisting of $K d_{1 \times M}$ kernels. A cell (red-dashed box) contains a tunable power splitter for power distribution and routing, a PCM memory

(or weight) for multiplication, a directional coupler for accumulation and a crossing for interconnect (Fig. 2c). The system scalability is evident from the periodic cell layout in the 2D plane. MVM requires equal power distribution to all the PCM weights and the same contribution from different cells for linear accumulation. The requirements are fulfilled by a meticulous design of the power splitter and directional coupler (Supplementary Section 1). In addition to equal power distribution, power splitters also serve to concentrate all the optical power in a specific cell during the PCM weight-setting process (Methods). The input data architecture features a 2D array

$$X_{2D} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \dots & x_{MN} \end{bmatrix}$$

representing multiple $d_{M \times 1}$ vectors. Here N RF components are multiplexed to produce this $d_{M \times N}$ matrix. The n th vector is carried by the corresponding RF component at frequency f_n . Data in the m th row are carried by a continuous-time signal $in_m(t) = \sum_{n=1}^N x_{mn} e^{i2\pi f_n t}$ through encoding individual elements into amplitudes of N different RF components and input via optical channel m of the photonic tensor core. The weighted sum of M such inputs that is output from column k is

$$out(t)_k = \sum_{m=1}^M w_{km} in_m(t) = \sum_{n=1}^N \sum_{m=1}^M w_{km} x_{mn} e^{i2\pi f_n t},$$

whose Fourier transform is $out(f)_k = \sum_{n=1}^N \sum_{m=1}^M w_{km} x_{mn} \delta(f - f_n)$. Consequently, the collective outputs from all the columns are

$$Y = \begin{bmatrix} out(f)_1 \\ out(f)_2 \\ \vdots \\ out(f)_K \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^N \sum_{m=1}^M w_{1m} x_{mn} \delta(f - f_n) \\ \sum_{n=1}^N \sum_{m=1}^M w_{2m} x_{mn} \delta(f - f_n) \\ \vdots \\ \sum_{n=1}^N \sum_{m=1}^M w_{Km} x_{mn} \delta(f - f_n) \end{bmatrix} = \begin{bmatrix} y_{11}y_{12} \dots y_{1N} \\ y_{21}y_{22} \dots y_{2N} \\ \vdots \\ y_{K1}y_{K2} \dots y_{KN} \end{bmatrix}$$

Y represents N MVM results of all the $N d_{M \times 1}$ vectors in X_{2D} multiplied by the kernel matrix W . Using Q WDM channels will result in $Q \times N$ MVMs.

Verification of fundamental operations

The additional RF DOF is introduced to the system using a continuous-time data representation. We first verify the feasibility of using continuous-time data representation for photonic in-memory computing. A photonic tensor core provides three fundamental functions: data summation by routing cell outputs to common buses, data weighting by PCM memory and consequent weighted data summation. These three functions correspond to three mathematical operations, namely, addition, multiplication and multiply–accumulate (MAC), respectively. These three operations are studied using a Y junction loaded with PCM memories on both arms (Fig. 3a). Supplementary Section 2 shows the representative scanning electron microscopy image and testing setups. Fifty RF components ($N = 50$) are multiplexed to generate $d_{1 \times 50}$ input vectors. The frequencies of these 50 RF components uniformly span from $f_1 = 0.15$ MHz to $f_{50} = 2.60$ MHz. The shortest acquisition time required is $t_{min} = 1/\text{gcd}(f_1, f_2, \dots, f_{50})$, such that an integer multiple of complete waveforms can be acquired for each RF component, where gcd stands for

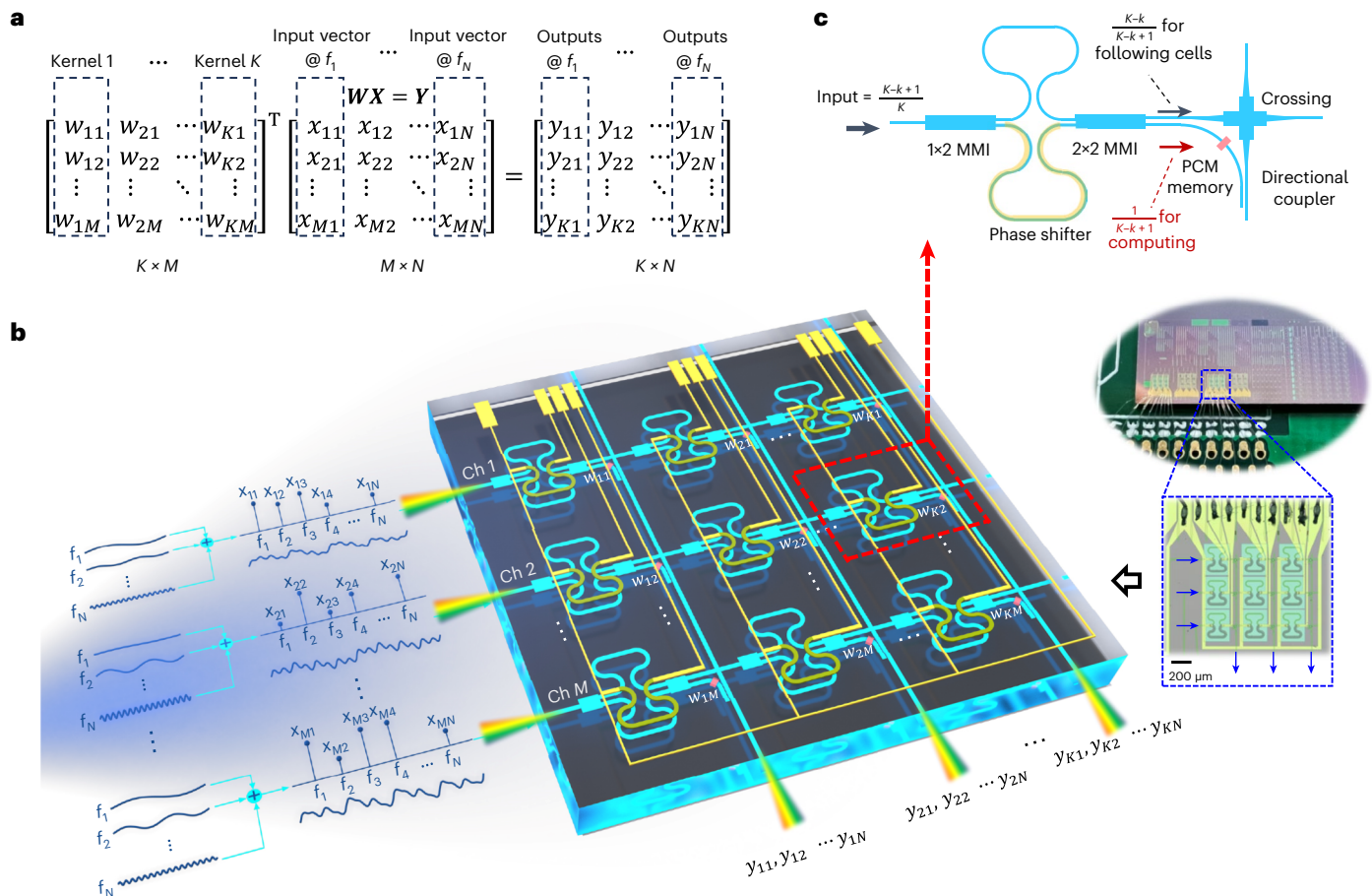


Fig. 2 | Data architecture and working principle of a photonic tensor core for in-memory computing using continuous-time data representation. **a**, Target matrix–matrix multiplication using only one optical wavelength and N multiplexed RF components. **b**, Implementation of the matrix–matrix multiplication. The weight matrix W of dimension $d_{K \times M}$ containing K $d_{1 \times M}$ kernels is defined by the tensor core with M inputs and K outputs. Carried by one wavelength λ_L , a $d_{M \times N}$ matrix X is input using M input optical channels and N multiplexed RFs. The n th $d_{M \times 1}$ vector $(x_{1n}, x_{2n}, \dots, x_{Mn})^T$ is encoded in the amplitude

of RF f_n . The m th element is input via input optical channel m . Consequently, Q matrix–matrix multiplications can be processed in parallel using Q wavelengths, where each wavelength carries a $d_{M \times N}$ matrix. **c**, Each cell (highlighted in the red-dashed box in **b**) in the tensor core contains a tunable power splitter for optical power distribution and routing, a PCM memory for multiplication, a waveguide crossing for interconnect and a directional coupler for addition. Here k represents the k th output column.

the greatest common divisor. All the numbers are randomly generated from $[0, 1]$ with a 0.01 resolution.

Supplementary Section 3 shows the basic transmission performance of multiplexed RF modulated optical signal. To verify the addition operation, two weights are idle. Each value in a $d_{1 \times 50}$ vector $(x_1, x_2, \dots, x_{50})$ is encoded in the respective RF amplitude. Two multiplexed RFs modulate two optical carriers to generate continuous-time inputs, namely, $in_1(t) = \sum_{j=1}^N x_j e^{i2\pi f_j t}$ and $in_2(t) = \sum_{k=1}^N x_k e^{i2\pi f_k t}$. The time-domain output is the direct sum of the two inputs, that is, $out(t) = \sum_{j=1}^N x_j e^{i2\pi f_j t} + \sum_{k=1}^N x_k e^{i2\pi f_k t}$ (Fig. 3b), and the frequency-domain output is the sum of two RF amplitudes at discrete frequencies, that is, $out(f) = \sum_k \sum_j (x_k + x_j) \delta(f_k - f_j)$ (Fig. 3c). The accuracy of the addition operation is revealed by its error distributions (Supplementary Section 4). The wavelength spacing ($\Delta\lambda$) between the two inputs is also studied (Supplementary Section 5) for harnessing dense WDM parallelism in system implementation. The accuracy of the addition operation under different numbers of multiplexed RF components is also studied (Supplementary Section 6), suggesting that $N = 50$ presented here is not a limitation of parallelism for low-precision ML models³⁸. To verify the multiplication operation, only one arm of the Y junction is active. A continuous-time input consisting of multiplicands is $in(t) = \sum_{j=1}^N x_j e^{i2\pi f_j t}$. The multiplier w (or weight) is determined by the crystalline state of PCM and can be set using optical pulses

(Supplementary Section 7). The resultant change in optical transmission

$$\Delta T = \frac{T_{\text{set}} - T_{\text{ref}}}{T_{\text{ref}}}$$

can be continuously tuned from 0% to more than 20% by increasing the amorphization pulse width (Fig. 3d). The weight w can be mapped to $[0, 1]$, leading to normalized outputs from PCM memory: $w \times x \in [0, 1]$. Supplementary Section 8 describes the details of weight mapping. The frequency-domain outputs at different weights are examined to confirm that the multiplicands encoded in the different RF components are operated by the same multiplier (Supplementary Section 9). The accuracy of the multiplication operation is revealed by the Gaussian error distribution of 1,500 multiplication results, obtained by multiplying 300 random numbers and 5 weights, showing a standard deviation of 0.056 ± 0.001 (Fig. 3e). The whole Y junction is active for the verification of the two-channel MAC operation. The input vectors and operation principle are similar to the combination of addition and multiplication operations. Using 300 pairs of random numbers and 5 pairs of weights using just a Y junction, we obtain a standard deviation of 0.057 ± 0.001 in Gaussian error distribution from 1,500 MAC operations (Fig. 3f). In a photonic tensor core with 300 three-element arrays of random numbers and 5 three-element arrays of weights, the standard deviation we record is 0.063 ± 0.001 (Supplementary Section 10), where the expected performance of using more optical channels is also estimated. The errors are

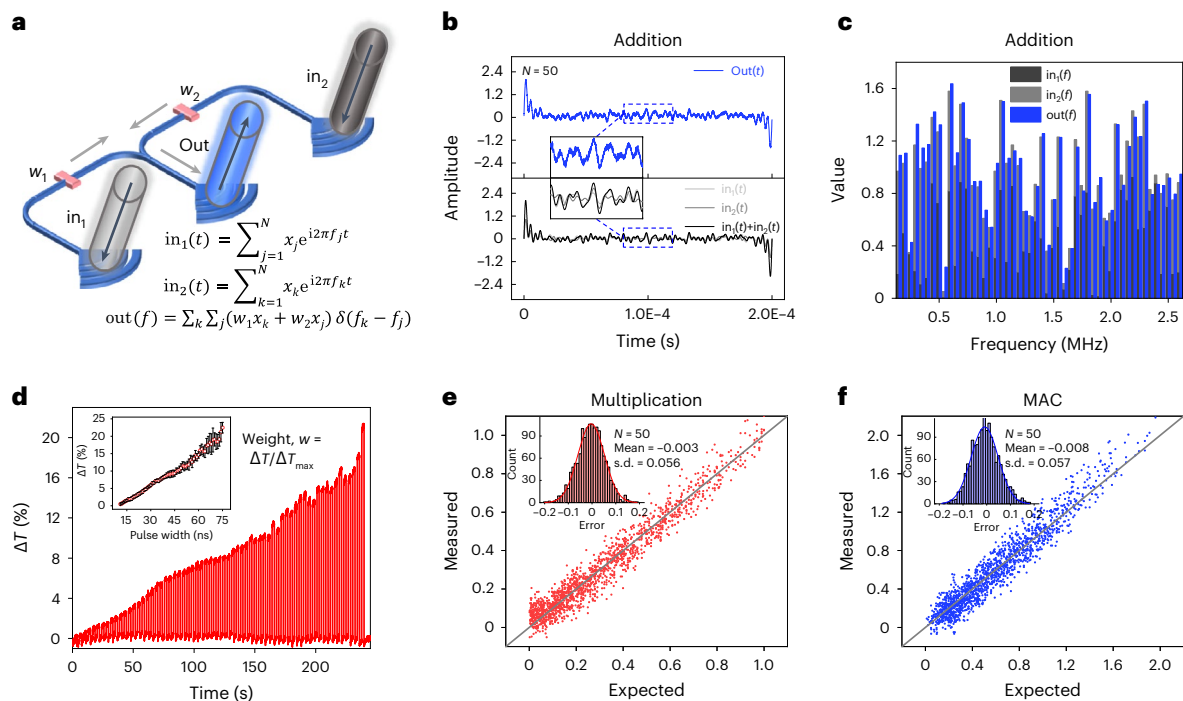


Fig. 3 | Photonic addition, multiplication and MAC operations using continuous-time data representation with 50 multiplexed RFs. All the input numbers x are randomly generated from $[0, 1] \subseteq R$ with a 0.01 resolution. **a**, Y junction loaded with PCM memories on each arm for the verification of operations. **b, c**, Comparison of normalized measured and expected time-domain addition output (**b**) and frequency-domain output (**c**). **d**, Quasi-analogue PCM

weight setting using optical pump pulses with varying widths. **e**, Accuracy of 1,500 multiplication results from 300 random multiplicands and 5 multipliers. The inset shows the normalized error distribution. s.d., standard deviation. **f**, Accuracy of 1,500 MAC results using 300 pairs of random numbers and 5 pairs of weights. The inset shows the normalized error distribution.

attributed to variations in the weight setting and noise from receivers. The former can be minimized by the progressive setting method that gradually increases the setting pulse energy until the desired transmission is reached³⁴, and the latter can be improved by using on-chip integrated photodetectors with a lower noise-equivalent power or reducing the optical loss of the PIC to enhance the signal-to-noise ratio.

This successful verification of three fundamental operations proves the feasibility of using a continuous-time data representation to add the RF DOF to photonic in-memory computing. Using multiplexed $N = 50$ RF components for a simple PCM-loaded Y junction, a parallelism of 50 is achieved, showing the high parallelism provided by the additional RF DOF. Importantly, this high parallelism contributed by RF can be conveniently incorporated into optoelectronic systems since it involves no additional optical multiplexing or filtering. Possible existing solutions to implement RF multiplexing include the use of field-programmable gate arrays and operational amplifier banks³⁹, making on-chip integration feasible for our proposed architecture.

Healthcare monitoring using a CNN

Statistics revealed by the World Health Organization show that CVDs are the leading cause of death, taking 17.9 million lives annually, with more than 80% caused by sudden heart attacks and strokes⁴⁰. Real-time ECG recording and analysis are crucial to minimize sudden death risks. An edge computing framework is a solution to simultaneously monitor the health condition of multiple CVD patients in real time with low latency⁴¹. The proposed computing architecture exploiting three DOFs is a potential platform to implement computing in edge clouds and perform the high-dimensional synchronous convolution of ECG signals and can facilitate ML-aided analysis to alert sudden death events, simultaneously benefiting a large number of CVD patients.

Having verified the feasibility of simultaneously using three DOFs, we configure our system to an architecture for edge cloud computing

(Fig. 4). Specifically, the wavelength and spatial DOFs are utilized for high-bandwidth parallel convolution and the RF DOF enables low latency and synchronization between the end devices. The system contains three layers (edge device, edge interface and edge cloud) with five functional blocks: input light generation and (de)multiplexing in the edge cloud, input-multiplexed RF generation at the edge device and interface, optical modulation relating edge interface and edge cloud, photonic tensor core for in-memory computing in the edge cloud, and output light (de)multiplexing and detection in the edge cloud. In our system implementation, six wavelengths covering 1,548.51 to 1,552.52 nm, with an adjacent spacing of 0.8 nm (equivalent to 100 GHz), are used for WDM. The highest RF frequency limited by our variable optical attenuators is 1 kHz. Methods and Supplementary Section 11 show the detailed system setup and electro-optic response, respectively. The corresponding operation is a specific case of the generalized data architecture and working principle described previously and is discussed in detail in Supplementary Section 12. In a single operation cycle, the system is synchronously performing 300 convolutions, convolving 100 ECG signals using three kernels.

The convolution results are further fed to a CNN for ML-aided ECG signal analysis. The CNN is designed to identify CVD patients at sudden death risks caused by ventricular fibrillation (a type of abnormal heart rhythm). The CNN architecture is illustrated with a single ECG signal without the loss of generality (Fig. 5a) and described in detail in Methods. Figure 5b shows a typical expected (Fig. 5b(i), convolved by CPU) and measured (Fig. 5b(ii), convolved by photonic system) convolution result of normal ECG signals, whereas Fig. 5c shows those in sudden death events. All the convolutions are performed once, and the error bands shown in Fig. 5b, c represent the standard deviation of convolution results from 50 pulses generated by the same patient, showing the variation in the ECG signal generated by this patient. Supplementary Figs. 15 and 16 show all the other convolution results.

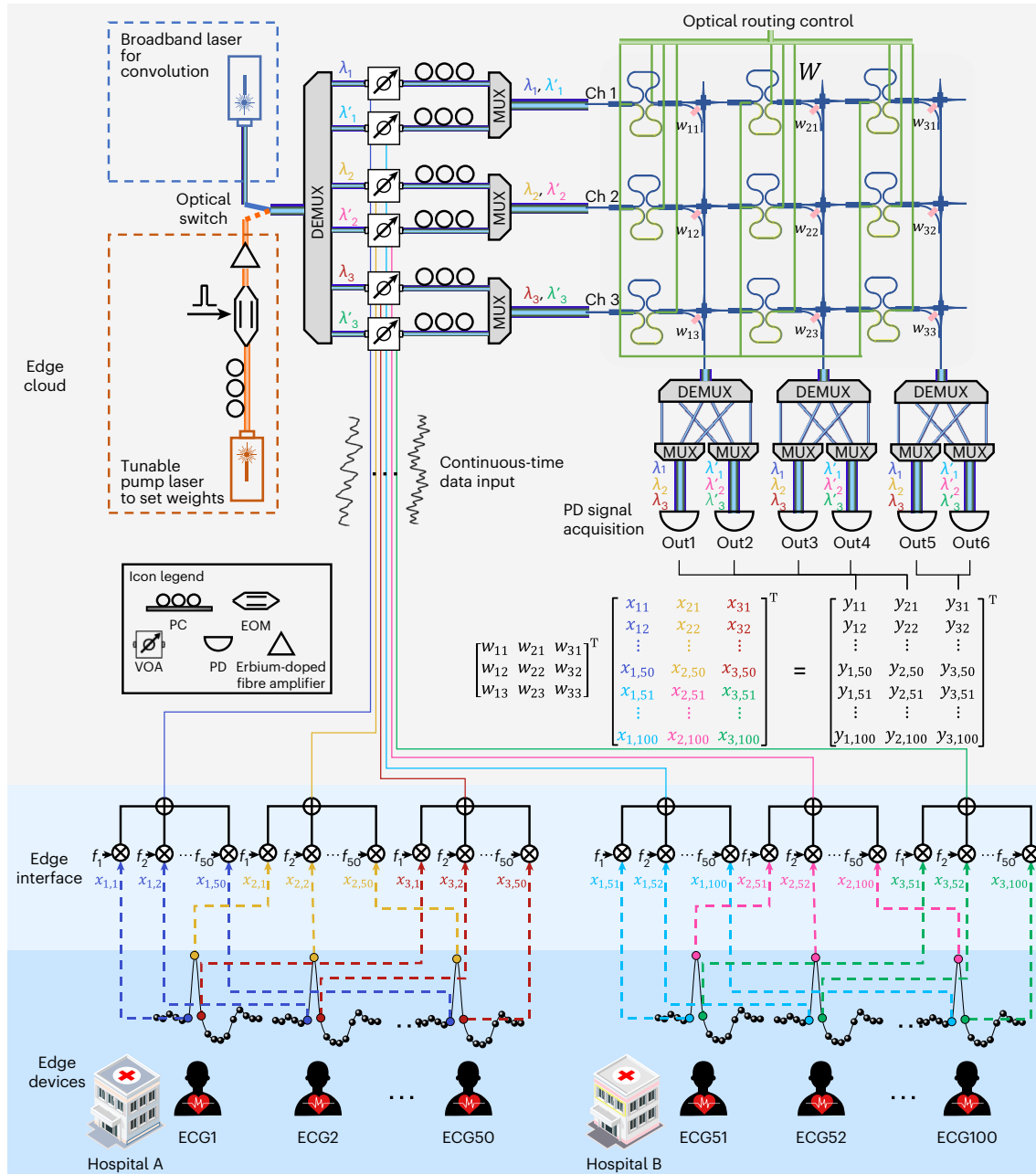


Fig. 4 | System architecture for edge cloud computing to synchronously convolve 100 clinical ECG signals from patients with CVD. The system has five functional blocks: input light generation and (de)multiplexing in the edge cloud, input-multiplexed RF generation at the edge device and interface, optical modulation relating edge interface and edge cloud, photonic tensor core for in-memory computing in the edge cloud, and output light (de)multiplexing and detection in the edge cloud. In the device layer, each ECG signal is a 1D time-domain signal. In the edge interface layer, the ECG signal data from patient j at time i are denoted as x_{ij} and encoded in the amplitude of RF $f_{\text{mod}(j,50)}$ using λ_i or

λ'_i as the carrier (λ_i if $j \leq 50$; λ'_i if $j > 50$). For $j \in [1, 100] \subseteq \mathbb{Z}^+$ and $i \in [1, 3] \subseteq \mathbb{Z}^+$, the input matrix X has dimension $d_{3 \times 100}$. In the edge cloud layer, the weight bank determined by the photonic tensor core defines a $d_{3 \times 3}$ matrix W , containing three $d_{1 \times 3}$ kernels. Effectively, one such matrix-matrix multiplication performs 300 convolutions resulting in a $d_{3 \times 100}$ matrix Y , which is obtained by convolving the middle three time-domain data of 100 ECG signals using 3 kernels. PD, photodetector; EOM, electro-optic modulator; PC, polarization controller; VOA, variable optical attenuator.

The features are effectively extracted, and the measured results resemble the expected ones. The convolution accuracy is examined by comparing 24,750 pairs of expected and measured results, showing a Gaussian error distribution with a low standard deviation of 0.015 ± 0.001 (Fig. 5d). Supplementary Fig. 17 shows the expected convolution result density. The standard deviation is lower than that obtained in MAC verification because most convolution results are small, within the range of $[0, 0.5]$. The CNN classification accuracies are presented in Fig. 5e. In the absence of a convolution layer, only

89% accuracy can be reached. With a convolution layer that helps to extract features, the accuracy is increased to 94.0% and 93.5% when the expected and measured convolution results are used, respectively. Minor differences in loss and accuracy evolution curves are observed between the use of expected and measured convolution results (Supplementary Fig. 18), suggesting a high accuracy of photonic-system-implemented convolution using continuous-time data representation. The confusion maps of classification results are shown in Supplementary Fig. 19, showing that there is only a 1%

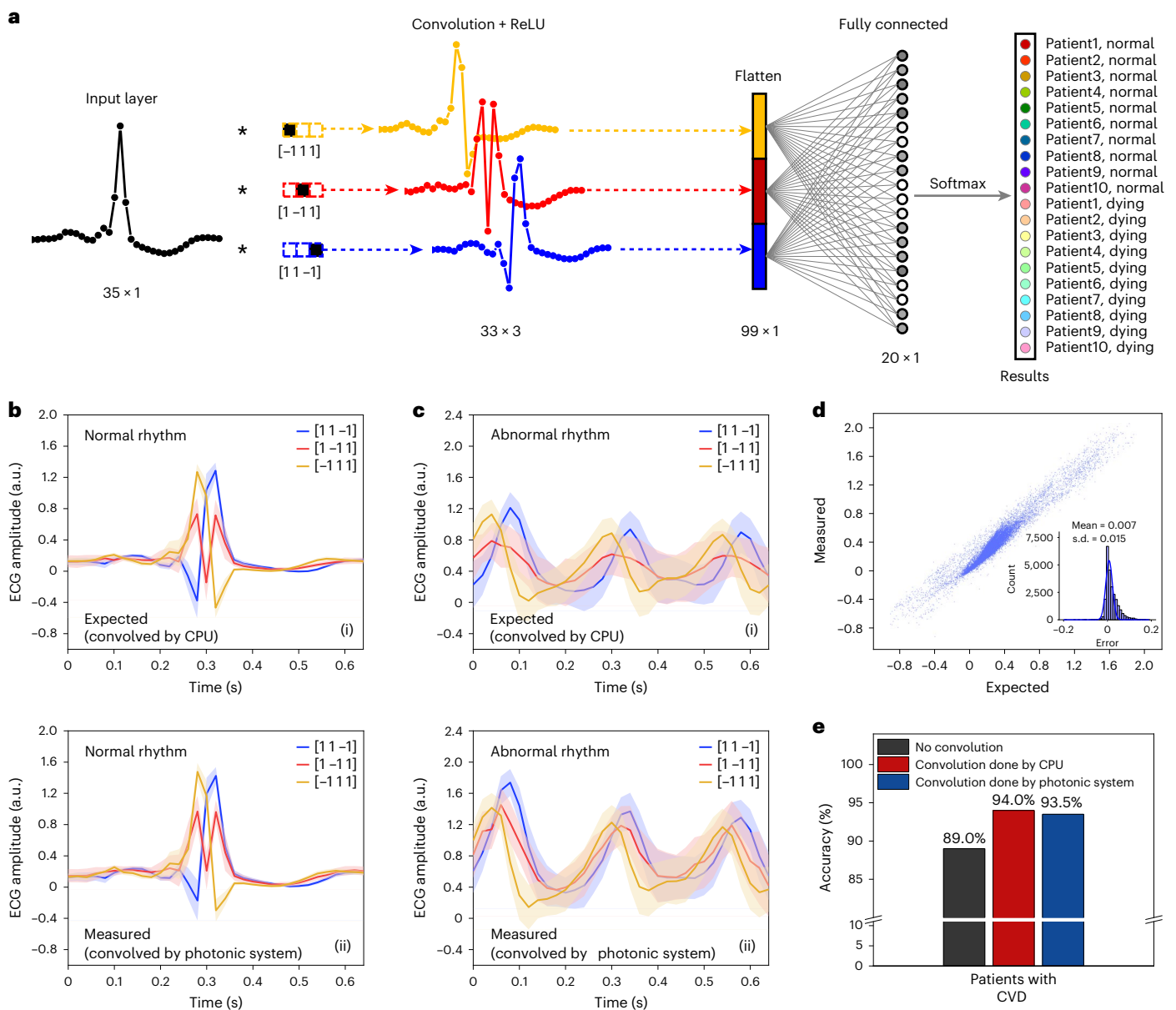


Fig. 5 | Healthcare monitoring of CVD patients using a CNN. **a**, CNN architecture. The CNN is designed to identify CVD patients at the risk of sudden death. ECG signals are supplied to the input layer. The system presented in Fig. 4 performs higher-dimensional convolution. A rectified linear unit (ReLU) layer, a fully connected layer and a softmax layer are applied in sequence after convolution. **b, c**, Comparison of expected convolution results (CPU convolved) (i) and measured convolution results (photonic system convolved) (ii) of normal ECG signals when patients are safe (**b**) and when patients are at risk when

experiencing ventricular fibrillation (**c**). All the convolutions are performed once, and the error bands represent the standard deviation of convolution results from 50 pulses generated by the same patient, showing the variation in the ECG signal generated by this patient. **d**, Convolution result accuracy. The inset shows the normalized error distribution. **e**, CNN classification accuracy. The classification accuracy using the measured convolution results (93.5%) is close to that using the expected convolution results (94.0%). Both accuracies are higher than that using the same neural network but without a convolution layer.

probability that abnormal ECG signals will be misclassified as normal ECG signals. Similar details are observed in the two maps, indicating the simultaneous achievement of high accuracy, effectiveness and ultraparallelism using our system that exploits three DOFs.

Discussion and conclusion

We have demonstrated the first instance of a photonic in-memory computing architecture capable of implementing higher-dimensional MVM in a single operation cycle of a physical device by increasing the multiplexing dimensionality using RF as a carrier. By verifying the feasibility of computing with continuous-time data in the optical domain, we provide an additional pathway to increase parallelism to photonic

processors. An electro-optically controlled photonic tensor core system was built to simultaneously exploit spatial, wavelength and RF DOFs to harness ultrahigh parallelism. A parallelism of 100, two orders higher than the previous implementation³⁴, was achieved by multiplexing 50 RF components on top of 2 WDM channels. Leveraging this higher-dimensional processing capability and high parallelism, we configured our system to an architecture for edge cloud computing to perform a synchronous convolution of 100 clinical ECG signals from CVD patients and built a CNN capable of identifying patients at sudden death risk with 93.5% accuracy. Although these are achieved using a small-size 3×3 photonic tensor core, larger-size photonic tensor cores are envisioned for better compute density, compute efficiency and

more general applications⁴². The scalability and performance estimation of larger-size photonic tensor cores are also discussed in detail (Supplementary Section 15). Crucially, the parallelism of 100 is not an upper limit (Supplementary Fig. 9); multiplexing 150 RF components is possible if lower precision is allowed. By using 16 WDM channels, an overall parallelism of 2,400 can also be achieved, suggesting that a single system can synchronously process signals from 2,400 end devices; this is currently not possible using existing technologies with lower-dimensional processing capability. Possible alternative methods towards this high computing capability include increasing the clock speed of electronics and using ultradense WDM channels. Supplementary Section 16 discusses the challenges associated with these two alternatives. Our proposed architecture is ubiquitously applicable to other photonic processing systems^{43–45} to enrich data information by exploiting more DOFs.

A key understanding underlying the mechanism of higher-dimensional data processing is that although the wavelength spacing of 0.8 nm may be considered ‘dense’ in WDM, this is orders of magnitude larger from an RF perspective. Therefore, the RF dimension can be regarded as a quasi-independent dimension that enriches data information. Meanwhile, continuous-time data representation brings another key advantage of avoiding electronic logic-state flips to potentially increase the clock frequency⁴⁶. More interestingly, the recent exploration of synthetic dimensions in photonics suggests that a single photonic cavity acousto-optic modulator naturally compatible with RF could be adopted to substantially reduce the footprint of the weighting matrix^{47,48}. From the hardware perspective, even though off-chip light sources, circulators, amplifiers, modulators and photodetectors were used in a lab environment aiming to verify high parallelism, these active photonic components can be monolithically integrated on a single chip^{29,49,50}. Complementary metal–oxide–semiconductor RF electronics can be adopted in the system to maximize the compute efficiency and density (Supplementary Section 17). In addition to the RF DOF, phase⁵¹, polarization⁵² and mode⁵³ DOFs of light could also offer more dimensions to further parallelize signal processing. However, the possible parallelism from these dimensions is restricted by their limited number of possible states and the requirement of waveguide compactness. It is also worth highlighting that the realization of ultrahigh parallelism relies on the combination of photonics that provides the wavelength DOF and electronics that provides the additional RF DOF, suggesting that synergy between photonics and electronics should be sought to fully unleash the potential of both in a single integrated system.

Online content

Any methods, additional references, Nature Portfolio 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/s41566-023-01313-x>.

References

- Statista Research Department. Amount of data created, consumed, and stored 2010–2020, with forecasts to 2025. Statista <https://www.statista.com/statistics/871513/worldwide-data-created/> (2022).
- Zhou, L., Pan, S., Wang, J. & Vasilakos, A. V. Machine learning on big data: opportunities and challenges. *Neurocomputing* **237**, 350–361 (2017).
- Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K. & O’Doherty, J. P. Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nat. Hum. Behav.* **5**, 743–755 (2021).
- Han, C. et al. Speaker-independent auditory attention decoding without access to clean speech sources. *Sci. Adv.* **5**, eaav6134 (2019).
- Assael, Y. et al. Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022).
- Rao, Z. et al. Machine learning-enabled high-entropy alloy discovery. *Science* **85**, 78–85 (2022).
- Fawzi, A. et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53 (2022).
- Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **56**, 49–56 (2022).
- Reuther, A. et al. AI accelerator survey and trends. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)* 1–9 (IEEE, 2021).
- Li, X., Zhang, G., Huang, H. H., Wang, Z. & Zheng, W. Performance analysis of GPU-based convolutional neural networks. In *Proc. International Conference on Parallel Processing* 67–76 (IEEE, 2016).
- Wang, Y. E., Wei, G.-Y. & Brooks, D. Benchmarking TPU, GPU, and CPU platforms for deep learning. Preprint at <http://arxiv.org/abs/1907.10701> (2019).
- Wang, L. et al. Superneurons: dynamic GPU memory management for training deep neural networks. *ACM SIGPLAN Not.* **53**, 41–53 (2018).
- Qiu, J., Wang, J., Yao, S., Guo, K. & Li, B. Going deeper with embedded FPGA platform for convolutional neural network. In *Proc. 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* 26–35 (ACM, 2016).
- Magaki, I., Khazraee, M., Gutierrez, L. V. & Taylor, M. B. ASIC clouds: specializing the datacenter. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* 178–190 (IEEE, 2016).
- Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
- Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
- Lanza, M. et al. Memristive technologies for data storage, computation, encryption, and radio-frequency communication. *Science* **376**, eabj9979 (2022).
- Wan, W. et al. A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504–512 (2022).
- Sarwat, S. G., Kersting, B., Moraitis, T., Jonnalagadda, V. P. & Sebastian, A. Phase-change memtransistive synapses for mixed-plasticity neural computations. *Nat. Nanotechnol.* **17**, 507–513 (2022).
- Kim, M. K., Kim, I. J. & Lee, J. S. CMOS-compatible compute-in-memory accelerators based on integrated ferroelectric synaptic arrays for convolution neural networks. *Sci. Adv.* **8**, eabm8537 (2022).
- Jung, S. et al. A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211–216 (2022).
- Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
- Zhou, H. et al. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light: Sci. Appl.* **11**, 30 (2022).
- Nahmias, M. A. et al. Photonic multiply-accumulate operations for neural networks. *IEEE J. Sel. Topics Quantum Electron.* **26**, 7701518 (2020).
- Yan, T. et al. All-optical graph representation learning using integrated diffractive photonic computing units. *Sci. Adv.* **8**, eabn7630 (2022).
- Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **15**, 102–114 (2021).
- Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).

29. Shu, H. et al. Microcomb-driven silicon photonic systems. *Nature* **605**, 457–463 (2022).
30. Tran, M. A. et al. Extending the spectrum of fully integrated photonics to submicrometre wavelengths. *Nature* **610**, 54–60 (2022).
31. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
32. Tait, A. N., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Broadcast and weight: an integrated network for scalable photonic spike processing. *J. Light. Technol.* **32**, 4029–4041 (2014).
33. Xu, X. et al. 11TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
34. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
35. Sludds, A. et al. Delocalized photonic deep learning on the Internet's edge. *Science* **378**, 270–276 (2022).
36. Rios, C. et al. Integrated all-photonic non-volatile multi-level memory. *Nat. Photon.* **9**, 725–732 (2015).
37. Wang, C. et al. Scalable massively parallel computing using continuous-time data representation in nanoscale crossbar array. *Nat. Nanotechnol.* **16**, 1079–1085 (2021).
38. Wang, N., Chen, C. & Gopalakrishnan, K. Ultra-low precision 4-bit training of deep neural networks. In *NIPS'20: Proc. 34th International Conference on Neural Information Processing Systems* 1796–1807 (IEEE, 2020).
39. Baig, M. T. et al. A scalable, fast, and multichannel arbitrary waveform generator. *Rev. Sci. Instrum.* **84**, 124701 (2013).
40. World Health Organization. Cardiovascular diseases; https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
41. Shi, W., Cao, J., Member, S. & Zhang, Q., Member, S. Edge computing: vision and challenges. *IEEE Internet Things J.* **3**, 637–646 (2016).
42. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
43. Dong, B. et al. Biometrics-protected optical communication enabled by deep learning-enhanced triboelectric/photonic synergistic interface. *Sci. Adv.* **8**, eabl9874 (2022).
44. Wu, C. et al. Harnessing optoelectronic noises in a photonic generative network. *Sci. Adv.* **8**, eabm2956 (2022).
45. Liu, W. et al. A fully reconfigurable photonic integrated signal processor. *Nat. Photon.* **10**, 190–195 (2016).
46. Markov, I. L. Limits on fundamental limits to computation. *Nature* **512**, 147–154 (2014).
47. Zhao, H., Li, B., Li, H. & Li, M. Enabling scalable optical computing in synthetic frequency dimension using integrated cavity acousto-optics. *Nat. Commun.* **13**, 5426 (2022).
48. Yuan, L., Lin, Q., Xiao, M. & Fan, S. Synthetic dimension in photonics. *Optica* **5**, 1396–1405 (2018).
49. White, A. D. et al. Integrated passive nonlinear optical isolators. *Nat. Photon.* **17**, 143–149 (2022).
50. Liu, Y. et al. A photonic integrated circuit-based erbium-doped amplifier. *Science* **376**, 1309–1313 (2022).
51. Ji, H. et al. 1.28-Tb/s demultiplexing of an OTDM DPSK data signal using a silicon waveguide. *IEEE Photon. Technol. Lett.* **22**, 1762–1764 (2010).
52. Lee, J. S., Farmakidis, N., Wright, C. D. & Bhaskaran, H. Polarization-selective reconfigurability in hybridized-active-dielectric nanowires. *Sci. Adv.* **8**, eabn9459 (2022).
53. Yang, K. Y. et al. Multi-dimensional data transmission using inverse-designed silicon photonics and microcombs. *Nat. Commun.* **13**, 7862 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Methods

Device fabrication

Waveguide devices for verification of basic operations. The fabrication started from a silicon-on-insulator wafer (Soitec) with a 220 nm silicon (Si) device layer and a 2 μm buried oxide layer. A 400-nm-thick positive electron-beam resist (CSAR-62) was spin coated on a diced 1 cm \times 1 cm silicon-on-insulator chip, followed by 3 min of pre-baking at 150 $^{\circ}\text{C}$. The electron-beam resist was patterned by electron-beam lithography (EBL; JEOL JBX-5500 50 kV) and developed in AR600-546 for 30 s, methyl isobutyl ketone for 15 s and isopropanol for 15 s in sequence. The waveguide patterns were transferred to the Si device layer (etch depth, 110 nm) by reactive ion etching (Oxford Instrument PlasmaPro) with SF_6 and CHF_3 gases, followed by O_2 plasma cleaning of CSAR. Next, a 2- μm -thick double-layer PMMA (PMMA 495 A8 and PMMA 950 A8) was spin coated on the chip, followed by EBL patterning and development in methyl isobutyl ketone:isopropanol = 1:3 for 1 min to define the sputtering windows. A 10-nm-thick/10-nm-thick $\text{Ge}_2\text{Sb}_2\text{Te}_3$ (GST)/indium tin oxide (ITO) stack was deposited on the waveguide using a magnetron sputtering system (PVD, AJA International). The GST and ITO targets were sputtered at 30 W RF power with 3 s.c.c.m. Ar flow and 40 W RF power with 3 s.c.c.m. Ar flow, respectively, at a base pressure of 10^{-7} torr. The stack was then lifted off in acetone for 180 min at 50 $^{\circ}\text{C}$. Finally, the chip was annealed on a hotplate for 5 min at 250 $^{\circ}\text{C}$ to fully crystallize the GST.

Electro-optically controlled photonic tensor core. The passive silicon photonic circuit was fabricated using the foundry multi-project wafer service provided by CORNERSTONE. The detailed specifications of CORNERSTONE standard waveguide components can be found at <https://cornerstone.sotonfab.co.uk/>. The fabricated Si photonic circuit has a 1- μm -thick silicon dioxide (SiO_2) upper cladding. SiO_2 windows were patterned by EBL and opened by hydrogen fluoride for the subsequent deposition of the GST/ITO stack, which is similar to the previously described GST/ITO sputtering procedure. Next, NiCr heater patterns were defined by EBL using a double-layer PMMA (PMMA 495-A3 and PMMA 495-A6) as the photoresist. A 200-nm-thick NiCr layer was sputtered followed by PMMA lift-off to form NiCr heaters. Gold pads with 75 nm thickness were fabricated using a similar process as the NiCr heater fabrication, but with thermal evaporation (Edwards 306). A 3–5 nm Cr layer is deposited before gold deposition to serve as an adhesion layer. The chip was then annealed on a hotplate for 5 min at 250 $^{\circ}\text{C}$ to fully crystallize the GST. Finally, the chip was wire bonded to a printed circuit board for electro-optic control.

Measurement setup

Setup for verification of operations using continuous-time data representation. Supplementary Section 2 comprehensively describes the experimental setups used to verify the fundamental operations using continuous-time data representation. The setup used to verify the transmission operation and the multiplication operation is an optical waveguide pump–probe setup (Supplementary Fig. 3), which was reported before⁵⁴. The pump line and probe line were taking opposite routes in the waveguide by using two fibre-optic circulators. The full setup was used for multiplication. The pump laser line was idle in transmission. The setup used to verify the addition and MAC operations is a modified optical waveguide pump–probe setup that accommodates a Y junction (Supplementary Fig. 4). The pump line and probe line followed the same route in the waveguide. The full setup was used for verifying the MAC operation. The pump laser line was idle in verifying the addition operation.

System setup for synchronous convolution. The experimental setup for the synchronous convolution of 100 ECG signals is shown in Fig. 4. The photonic tensor core has three input optical channels and three output optical channels, representing a $d_{3 \times 3}$ matrix consisting of three $d_{1 \times 3}$

kernels. The input light was switchable between a supercontinuum laser (SuperK COMPACT, NKT Photonics) and a tunable pump laser (Santec, TSL-550) using an optical switch (Gezhi GZ-12C-1 \times 2-SM). The PCM memory in each cell of the photonic tensor core was first set to the desired weight to correctly define the kernels. The tunable pump laser was used in the PCM weight setting. The amplified pump light passed through a demultiplexer (DEMUX) module (Gezhi, DWDM-100G-DEMUX) so that different wavelengths were routed to different input optical channels ($\lambda_1 = 1,552.52$ nm to optical channel 1, $\lambda_2 = 1,551.72$ nm to optical channel 2 and $\lambda_3 = 1,550.92$ nm to optical channel 3). The tunable power splitters of the photonic tensor core were controlled by a microprocessing unit (Analog Devices DC2026) to ensure that all the pump power was concentrated into the PCM of the target cell. For example, to set w_{23} , λ_3 was used so that the pump light was routed to optical channel 3. Cell₁₃ was controlled to distribute all the light into the top channel of its 2 \times 2 multimode interferometer (MMI), and cell₂₃ was controlled to distribute all the light into the bottom channel of the MMI to efficiently set w_{23} . In this case, cell₃₃ was idle. After setting all the PCM weights, a parallel convolution was performed using the supercontinuum laser. The DEMUX module was used to separate six wavelengths with a spacing of 0.8 nm (equivalent to 100 GHz) to different optical channels ($\lambda_1 = 1,552.52$ nm, $\lambda_2 = 1,551.72$ nm, $\lambda_3 = 1,550.92$ nm, $\lambda'_1 = 1,550.12$ nm, $\lambda'_2 = 1,549.32$ nm and $\lambda'_3 = 1,548.51$ nm). The ECG signal data were loaded onto each wavelength using a variable optical attenuator (VOA; Thorlabs V1550A). The VOAs with the highest RF frequency of 1 kHz were driven via coaxial cables by a digital signal processor (NI USB-6259) that generated 50 multiplexed RF components. Note that in practice, when the RF frequency is high (in the gigahertz range) and the transmission distance is long (>10 m) in the edge cloud computing framework, coaxial cables should be replaced by fibre-optic connections to avoid the power loss of high-frequency signals in the coaxial cables. Here λ_1 to λ_3 were carrying three respective time-domain data points of the ECG signals 1–50, whereas λ'_1 to λ'_3 were carrying the same data of ECG signals 51–100. The polarization of output light from VOA was controlled by a polarization controller (Thorlabs FPC032). The six wavelengths were then grouped by a multiplexer (MUX) array (Gezhi, DWDM-100G-MUX) to form three inputs to the respective input optical channels of the photonic tensor core (λ_1 and λ'_1 to optical channel 1, λ_2 and λ'_2 to optical channel 2 and λ_3 and λ'_3 to optical channel 3). Convolutions were naturally performed as light propagated through the photonic tensor core. Each output optical channel of the photonic tensor core contained all the wavelengths λ_1 – λ_3 and λ'_1 – λ'_3 . These six wavelengths were demultiplexed and regrouped by a MUX/DEMUX array to form two groups of multiplexed output. Here λ_1 – λ_3 formed one group representing the convolution results of three time-domain data points of ECG signals 1–50 and λ'_1 – λ'_3 formed another group representing the same representation but for ECG signals 51–100. The resultant six groups of output light were detected by a photodetector array (Newport New Focus 2011).

Generation, convolution and output of multiplexed RF signals

The properties of the original ECG data collected from Holter monitors are described in the ‘ECG signal dataset’ section. The Holter monitors represent the edge device layer. The generation of multiplexed RF signals represents the operations performed in the edge interface layer. The convolution and output are implemented in the edge cloud layer.

For parallel convolution of the middle three time-domain data of 100 ECG signals, the input matrix is a $d_{3 \times 100}$ matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,100} \\ x_{21} & x_{22} & \cdots & x_{2,100} \\ x_{31} & x_{32} & \cdots & x_{3,100} \end{bmatrix}.$$

The j th column of X contains the middle three time-domain data of the j th ECG signal (Fig. 4). The i th row of X contains the i th time-domain data of 100 ECG signals. Taking the first row ($x_{11}x_{12}\dots x_{1,100}$), for example,

the j th element x_{ij} , where $j \in [1, 100] \subseteq \mathbb{Z}^+$, was encoded in the amplitude of the RF component $f_{\text{mod}(j,50)}$, resulting in a continuous-time data representation of $x_{ij}e^{i2\pi f_j t}$. The whole row was represented by the multiplexed RF signal $\text{in}_1(t) = \sum_{j=1}^{100} x_{ij}e^{i2\pi f_j t}$. Similarly, $\text{in}_2(t) = \sum_{k=1}^{100} x_{2k}e^{i2\pi f_k t}$ and $\text{in}_3(t) = \sum_{l=1}^{100} x_{3l}e^{i2\pi f_l t}$. The three inputs with continuous-time data representation were mathematically generated in MATLAB R2021b, and converted to .tfw files⁵⁵ readable by a function generator (Tektronix AFG3102C). The subsequent electrical output from the function generator drove VOAs to load the ECG data into the optical domain. Here $\text{in}_1(t)$ to $\text{in}_3(t)$ were input to optical channel 1 to channel 3, respectively. The photonic tensor core was then effectively performing:

$$Y(t) = W \cdot X(t) = \begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \end{bmatrix}^T \begin{bmatrix} \sum_{j=1}^{100} x_{1j}e^{i2\pi f_j t} \\ \sum_{k=1}^{100} x_{2k}e^{i2\pi f_k t} \\ \sum_{l=1}^{100} x_{3l}e^{i2\pi f_l t} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{100} w_{11}x_{1j}e^{i2\pi f_j t} + \sum_{k=1}^{100} w_{21}x_{2k}e^{i2\pi f_k t} + \sum_{l=1}^{100} w_{31}x_{3l}e^{i2\pi f_l t} \\ \sum_{j=1}^{100} w_{12}x_{1j}e^{i2\pi f_j t} + \sum_{k=1}^{100} w_{22}x_{2k}e^{i2\pi f_k t} + \sum_{l=1}^{100} w_{32}x_{3l}e^{i2\pi f_l t} \\ \sum_{j=1}^{100} w_{13}x_{1j}e^{i2\pi f_j t} + \sum_{k=1}^{100} w_{23}x_{2k}e^{i2\pi f_k t} + \sum_{l=1}^{100} w_{33}x_{3l}e^{i2\pi f_l t} \end{bmatrix}$$

The frequency-domain representation of Y is

$$Y(f) = \begin{bmatrix} \sum_{l=1}^{100} \sum_{k=1}^{100} \sum_{j=1}^{100} (w_{11}x_{1j} + w_{21}x_{2k} + w_{31}x_{3l}) \times \delta(f_j - f_k)(f_k - f_l) \\ \sum_{l=1}^{100} \sum_{k=1}^{100} \sum_{j=1}^{100} (w_{12}x_{1j} + w_{22}x_{2k} + w_{32}x_{3l}) \times \delta(f_j - f_k)(f_k - f_l) \\ \sum_{l=1}^{100} \sum_{k=1}^{100} \sum_{j=1}^{100} (w_{13}x_{1j} + w_{23}x_{2k} + w_{33}x_{3l}) \times \delta(f_j - f_k)(f_k - f_l) \end{bmatrix}$$

$$= \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1,100} \\ y_{21} & y_{22} & \cdots & y_{2,100} \\ y_{31} & y_{32} & \cdots & y_{3,100} \end{bmatrix}$$

where $y_{ij} = w_{1i}x_{1j} + w_{2i}x_{2j} + w_{3i}x_{3j}$ was encoded in the RF component $f_{\text{mod}(j,50)}$, representing the convolution result of the middle three time-domain data of the j th ECG signal using the i th kernel. Each row of Y was output from the output optical channel of the respective photonic tensor core.

CNN model

ECG signal dataset. Long-time-duration ECG signals (shortest duration, 4 h 15 min 10 s) from ten CVD patients were taken from Sudden Cardiac Death Holter Database in PhysioNet^{56,57}. Supplementary Section 14 provides the corresponding clinical information of these ten patients. Here 50 normal pulses and 50 abnormal pulses were extracted from each patient, leading to a total of 500 normal pulses and 500 abnormal pulses. Each pulse has a 0.7 s duration. The original ECG signals have a 0.004 s time resolution. The ECG pulses were extracted with a time interval of 0.02 s (that is, one out of every five original dataset), leading to 35 datasets in the extracted ECG pulses. The 0.02 s time interval was carefully chosen to minimize the extracted dataset and maintaining the key features in the original ECG pulses. Here 80% of the pulses were used for training and 20% were used for testing, that is, a total of 800 pulses for training (400 normal pulses and 400 abnormal pulses) and 200 pulses for testing (100 normal pulses and 100 abnormal pulses).

CNN architecture. The CNN architecture is shown in Fig. 5a. The input layer takes the ECG pulse, which is in the form of a $d_{35 \times 1}$ 1D array. Time multiplexing is used to assist in sending the data of the ECG signals. At each time step, the convolution window takes three data points.

The window is moved by one data point after each step. Therefore, $35 - 3 + 1 = 33$ time steps are required to process the whole trace containing the 35 data points. This signal, represented as a 1D array is passed to a convolution layer consisting of three $d_{1 \times 3}$ kernels. Convolution operations were implemented with a stride of 1 and valid padding, resulting in a $d_{3 \times (35-3+1)}$ output. The output was activated by a rectified linear unit layer and flattened to a $d_{99 \times 1}$ vector. The flattened activated output was then fed to a fully connected layer with 20 neurons. The output from the fully connected layer was converted into probabilities by a softmax layer. Finally, the classification result was obtained. The ECG pulses were classified into 20 categories, representing two heart health conditions (normal or abnormal) of 10 individual patients. The convolution operations were implemented using the electro-optically controlled photonic tensor core system. The convolution results were processed by the following CNN layers using the deep learning toolbox in MATLAB R2021b. Weights of the fully connected layer were trained by the Adam optimizer. Here 100 epochs were used to reach the final CNN outcomes.

Data availability

The data that support the findings of this study are available from the corresponding author upon request. The ECG dataset analysed in this study is available from the open-source 'Sudden Cardiac Death Holter Database' via PhysioNet at <https://doi.org/10.13026/C2W306>. A sustainability report related to this article is available at <https://nanoeng.materials.ox.ac.uk/sustainability>.

Code availability

The code used in the present work is available from the corresponding author upon request.

References

- Ríos, C. et al. In-memory computing on a photonic platform. *Sci. Adv.* **5**, eaau5759 (2019).
- Sanz, M. createTFW(inputSignal, filename). *MATLAB Central File Exchange* (2022).
- Greenwald, S. D. *The Development and Analysis of a Ventricular Fibrillation Detector* (Massachusetts Institute of Technology, 1986).
- Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000).

Acknowledgements

H.B., C.D.W. and W.H.P.P. acknowledge support from the European Union's Horizon 2020 research and innovation programme (grant no. 101017237, PHOENICS Project). H.B. and W.H.P.P. acknowledge support from the European Union's Innovation Council Pathfinder programme (grant no. 101046878, HYBRAIN Project). B.D. acknowledges financial support from Singapore A*STAR International Fellowship (AIF).

Author contributions

All authors substantially contributed to this work. B.D. and H.B. conceived the experiment. B.D. and X.L. fabricated the devices with assistance from S.A., W.Z., N.F. and Y.H. B.D. implemented the measurement setup and carried out the measurements with help from S.A., W.Z., U.E.A., N.F., J.S.L. and Y.H. All authors discussed the data and wrote the manuscript together. H.B. led the work.

Competing interests

H.B. and W.H.P.P. hold shares in Saliency Labs Ltd. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41566-023-01313-x>.

Correspondence and requests for materials should be addressed to H. Bhaskaran.

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

Reprints and permissions information is available at www.nature.com/reprints.