Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Temporal speckle enables probabilistic weights for uncertainty-aware photonic AI

Wolfram Pernice

wolfram.pernice@kip.uni-heidelberg.de

Heidelberg University https://orcid.org/0000-0003-4569-4213

Frank Brückerhoff-Plückelmann

University of Heidelberg https://orcid.org/0009-0009-7713-6735

Hendrik Borras

Heidelberg University

Shivaprasad Umesh Hulyal

EPFL https://orcid.org/0009-0001-4675-9923

Lennart Meyer

Heidelberg University https://orcid.org/0009-0000-9381-9856

Xinru Ji

EPFL https://orcid.org/0000-0001-7057-7786

Jianqi Hu

École Polytechnique Fédérale de Lausanne https://orcid.org/0000-0002-2018-253X

Jiale Sun

Swiss Federal Institute of Technology Lausanne (EPFL)

Bernhard Klein

Heidelberg University https://orcid.org/0000-0003-0497-5748

Falk Ebert

Ruprecht-Karls-Univ Heidelberg

Jelle Dijkstra

Heidelberg University https://orcid.org/0009-0007-4647-4235

Liam McRae

Heidelberg University

Philipp Schmidt

Heidelberg University https://orcid.org/0009-0009-9290-3314

Tobias Kippenberg

Swiss Federal Institute of Technology Lausanne (EPFL) https://orcid.org/0000-0002-3408-886X

Holger Fröning

Heidelberg University https://orcid.org/0000-0001-9562-0680

Article

Keywords:

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10666437/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Temporal speckle enables probabilistic weights for uncertainty-aware photonic AI

F. Brückerhoff-Plückelmann[1*], H. Borras[2*], S. U. Hulyal[3*], L. Meyer[1], X. Ji[3], J. Hu[3], J. Sun[3], B. Klein[2], F. Ebert[1], J. Dijkstra[1], L. McRae[1], P. Schmidt[1], T. J. Kippenberg[3†], H. Fröning[2], W. Pernice [1†]

[1]*Kirchhoff-Institute for Physics, University of Heidelberg; Heidelberg, 69120, Germany.*

[2]*Institute of Computer Engineering, University of Heidelberg; Heidelberg, 69120, Germany.*

[3]*Institute of Physics, Swiss Federal Institute of Technology, Lausanne, CH-1015 Lausanne, Switzerland.*

[*]These authors contributed equally to this work.

[†]Corresponding author. Email: wolfram.pernice@kip.uni-heidelberg.de, tobias.kippenberg@epfl.ch

Artificial intelligence increasingly guides decisions in safety-critical settings, yet neural networks can remain overconfident when confronted with ambiguous observations or unfamiliar inputs. Bayesian neural networks quantify such uncertainty by replacing every weight with a probability distribution, but sampling many distributions at prediction time overwhelms deterministic digital hardware. Here we show that the temporal speckle of broadband light provides a native physical representation of probabilistic weights. Random beating among the frequency components of individually amplitude-bandwidth programmed spectral channels enables optical-encoding of mean and variance in each channel independently. This way, a single broadband light source stores many probabilistic weights in parallel and delivers mutually uncorrelated samples without requiring computationally demanding random-number generation. We harness this mechanism in an ultrafast photonic processor that computes one probabilistic convolution every 37.5 ps and allows simultaneous classification and out-of-distribution detection in hybrid Bayesian neural networks. Our system classifies blood-cell images while recognizing withheld cell types as unfamiliar, and separates epistemic novelty from aleatoric ambiguity on standard uncertainty benchmarks. Programming probability distributions into the fluctuations of light removes the sampling bottleneck of Bayesian inference and opens a route to high-speed, uncertainty-aware photonic AI.

Introduction

Artificial intelligence (AI) increasingly informs decisions in medicine, autonomous systems and other safety-critical settings, where an incorrect prediction can have serious consequences. Reliable deployment therefore requires models that recognize uncertainty arising from noisy or ambiguous observations (aleatoric uncertainty) and from inputs that differ from the training data (epistemic uncertainty) as illustrated in Fig. 1a. Conventional deterministic neural networks can remain confident in both situations and may return an incorrect prediction without warning [1, 2]. Bayesian neural networks address this limitation by inferring probability distributions over model parameters from the training data rather than single point estimates (Fig. 1b) [3, 4, 5, 6]. During prediction, repeated stochastic forward passes sample these distributions and produce a predictive distribution from which uncertainty can be quantified. This principled treatment of uncertainty, however, comes with substantial computational overhead from repeated sampling and probabilistic operations, limiting the use of Bayesian neural networks in complex, real-time applications. Addressing these limitations would not only accelerate AI, but also pave the way for a new generation of intelligent systems capable of reasoning under uncertainty instead of ignoring it.

The corresponding hardware challenge is not merely to generate random numbers, but to represent many independently programmable probability distributions and process their samples efficiently. Conventional digital accelerators generate and transfer stochastic weights between memory and compute units, increasing latency, energy consumption and memory traffic [7]. Analogue hardware can instead represent probabilistic parameters through intrinsic physical fluctuations and compute with them in situ. Such systems require stochastic signals whose statistics are independently tunable, stable and scalable across many weights. Electronic implementations based on memristive devices can control the mean and variance of stochastic conductances, but often require multiple devices per weight and remain affected by device variability, drift and limited bandwidth [8, 9, 10]. Photonics offers substantially larger bandwidth and sub-nanosecond processing [11], and previous photonic Bayesian approaches have explored programmed optical noise, chaotic light and interferometric stochastic devices [12, 13, 14]. What has remained missing is a physical mechanism that enables programmable weight distributions, parallel sampling and direct photonic computation while remaining compatible with established machine-learning frameworks.

Here we show that spectrally shaping broadband chaotic light provides such a mechanism. We encode each probabilistic weight in an individual frequency channel, using the integrated optical power to program its mean and the channel bandwidth to program its standard deviation (Fig. 1c). Combining this spectral encoding with high-speed temporal modulation and the frequency-dependent delay of an integrated chirped waveguide Bragg grating enables a photonic probabilistic convolution accelerator that computes one convolution output every 37.5 ps. A PyTorch-compatible interface connects the accelerator to deterministic digital layers and supports hardware-aware training of hybrid Bayesian neural networks. We experimentally deploy the resulting system for uncertainty-aware

blood-cell classification, detecting erythroblasts withheld from training as epistemically uncertain. We further evaluate the hybrid photonic network on MNIST, Fashion-MNIST and Ambiguous-MNIST, demonstrating the separation of epistemic novelty from aleatoric ambiguity [15]. These results establish spectrally encoded probabilistic weights as a route to high-speed uncertainty-aware photonic artificial intelligence.

Spectral encoding of probabilistic weights

We encode each probabilistic neural-network weight in a separate frequency channel of broadband chaotic light (Fig. 1b,c). The integrated channel power P_i programs the mean μ_i of weight i, whereas the optical bandwidth $\Delta\nu_i$ programs its standard deviation σ_i. Spectrally separated channels contain distinct frequency components and therefore produce mutually uncorrelated fluctuations [12]. A single broadband source can thus encode and sample multiple probabilistic weights in parallel. The bandwidth dependence originates from temporal averaging at the electronic readout, as sketched in Fig. 2a. Chaotic light with an optical bandwidth $\Delta\nu$ has a coherence time $\tau_c \sim 1/\Delta\nu$, whereas the finite electrical bandwidth B_{el} defines an effective measurement interval $\tau_m \sim 1/B_{el}$. For a narrow optical channel, the detector resolves the slowly varying power fluctuations. Increasing the optical bandwidth shortens the coherence time, so that each measurement averages over an increasing number of effective temporal coherence cells, $M_{eff} \sim \Delta\nu/B_{el}$. This averaging suppresses the relative fluctuations approximately as $M_{eff}^{-1/2}$. Optical power therefore sets the signal mean, whereas optical bandwidth controls the relative stochasticity.

We experimentally demonstrate this encoding using amplified spontaneous emission from an erbium-doped fibre source (Fig. 2b). A programmable waveshaper defines spectral channels with independently controlled power and bandwidth. We detect the resulting power fluctuations with a 38 GHz photodetector and record them using a real-time oscilloscope with a bandwidth of 33 GHz at 80 GSa s^{-1}. At fixed optical power, increasing the channel bandwidth progressively reduces the measured standard deviation because the detector averages over more temporal coherence cells (Fig. 2c). At large bandwidths, the standard deviation follows the expected $1/\sqrt{\Delta\nu}$ dependence before approaching the electronic-noise floor. At fixed bandwidth, increasing the optical power increases the absolute fluctuation amplitude approximately linearly. We extract M_{eff} from the measured voltage distributions at each operating point using a gamma distribution for the optical power fluctuations [16], a Gaussian distribution for the electronic ground noise and subtracting the mean signal due to the high-pass of the photodetector, also see the Supplementary Information[17]. The effective number of temporal coherence cells increases approximately linearly with optical bandwidth and remains largely independent of optical power as shown in Fig. 2d. The optical bandwidth also impacts the shape of the encoded distribution. When only a small number of coherence cells contributes to each measurement, the distribution retains the asymmetry of chaotic-light power fluctuations. Increasing the optical bandwidth averages over more independent fluctuations and produces a

narrower, more Gaussian-like distribution. Representative measurements in Fig. 2e demonstrate both control dimensions. At a fixed bandwidth of 50 GHz, increasing the optical power from 25 μW to 75 μW increases the standard deviation from 0.010 V to 0.029 V. At a fixed power of 75 μW, increasing the bandwidth from 50 GHz to 500 GHz narrows the distribution to a standard deviation of 0.012 V and makes it more Gaussian-like.

These measurements establish optical power and bandwidth as programmable physical parameters for encoding probabilistic weights. The chaotic optical field stores the distribution parameters and continuously generates new stochastic realizations within the same physical signal.

Photonic probabilistic convolution accelerator

We next combine spectral weight encoding with frequency–time interleaving [11] to compute convolutions with probabilistic weights directly in the optical domain, as illustrated in Fig. 3a. A waveshaper programs each weight distribution in a separate frequency channel of the amplified-spontaneous-emission spectrum. A broadband electro-optic modulator (EOM) encodes the same temporal input waveform onto all channels, and a chip-integrated chirped waveguide Bragg grating (CWBG) delays neighbouring channels by one input symbol. This frequency-dependent delay aligns each spectrally encoded weight with the corresponding input value at the photodetector. The detector sums the aligned optical powers, so that each measured output is a stochastic realization of the convolution between the input data and the programmed probabilistic kernel. The processor therefore generates the weight samples and evaluates the convolution within the same physical signal.

We experimentally implement a high-speed convolution accelerator using nine frequency channels centred around 194 THz and separated by 403 GHz. An 8-bit arbitrary-waveform generator (AWG) operating at 80 GSa s^{-1} drives the EOM and encodes each input value as a three-sample symbol with a duration of 37.5 ps. After modulation, a fibre polarization controller sets the input to the transverse-electric mode of the CWBG. An 8-bit real-time oscilloscope (RTO), also operating at 80 GSa s^{-1} and synchronized with the AWG, records the photodetector output. A PyTorch-compatible digital interface controls spectral programming, waveform generation and data acquisition. The integrated CWBG is fabricated on a low-loss silicon-nitride platform (Fig. 3b). Its measured frequency-dependent delay is $-93.12\,\mathrm{ps\,THz^{-1}}$ (Fig. 3e). Together with the 403 GHz channel spacing, this slope produces a delay of 37.5 ps between adjacent channels, exactly matching one encoded symbol. Once the pipeline is filled, the accelerator produces one probabilistic convolution output per symbol, corresponding to 26.7 billion outputs second. The measured delay from the AWG output to the RTO input remains approximately 70 ns across the operating band. Integrating the dispersive element on chip avoids the long fibre delay lines used in earlier deterministic frequency–time-interleaved processors [11] and provides the required channel alignment within a compact optical path and thus

with low latency.

We program all nine weight distributions simultaneously using iterative feedback. In each iteration, the accelerator evaluates a set of test convolutions from which we estimate the realized means and standard deviations. We then update the power and bandwidth of each frequency channel according to the difference between the measured and target distributions; the full procedure is described in the Supplementary Information [17]. This calibration programs the parameters of each distribution rather than individual stochastic samples. Once programmed, the chaotic optical field continuously generates new weight realizations without further spectral reconfiguration. Convolutional weight sharing also reuses the same programmed distributions across many input positions, amortizing the programming overhead over the subsequent high-speed operations.

We characterize the accelerator using 25 randomly generated probabilistic kernels. For each kernel, we evaluate a random input vector of length 1000 and acquire 1500 samples from every output distribution. The measured output means and standard deviations closely follow their target values across the accessible operating range, as shown in Fig. 3c,d. The errors, defined as measured minus target, have mean ± s.d. values of 0.016 ± 0.173 for the output mean and 0.036 ± 0.092 for the output standard deviation. To incorporate the photonic processor into Bayesian-network design, we represent it as a PyTorch-compatible probabilistic convolution operator. A differentiable physical model predicts the mean and standard deviation of the convolution output from the input data and programmed weight distributions, while calibrated stochastic perturbations capture the remaining deviations observed experimentally. This statistical abstraction avoids modelling every optical and electronic imperfection separately while retaining the processor behaviour relevant to probabilistic inference. We use the abstract operator during hardware-aware training and replace it with the physical accelerator during experimental prediction.

Uncertainty-aware photonic artificial intelligence

We co-design a Bayesian Neural Network to match the computational characteristics of the photonic Bayesian accelerator. The architecture contains a single probabilistic convolutional layer and combines DenseNet-style skip connections [18] with a modified form of the depthwise-separable convolutions introduced in MobileNetV1 [19], as shown in Fig. S2. An architecture search identifies the network structure and probabilistic-layer placement that maximizes the predictive performance, also see the Supplementary Information [17]. We implement the probabilistic convolution as a PyTorch-compatible operator [20, 21], allowing the deterministic digital layers to interact directly with either a software representation or the physical photonic accelerator. During training, a differentiable surrogate models the programmed weight distributions as Gaussian distributions and predicts the mean and standard deviation of the convolution output. We use stochastic variational inference (SVI) to infer these distributions from the training data [22]. Straight-through estimators

include non-differentiable hardware constraints, such as converter quantization, output clipping and calibrated computing errors, in the forward pass while preserving gradients during backpropagation. PyTorch and Pyro [23] provide automatic differentiation and probabilistic inference, implementation details are given in the Supplementary Information [17].

For experimental prediction, we replace the software surrogate with the physical accelerator. We evaluate each input in ten stochastic forward passes. In every pass, the accelerator evaluates the probabilistic convolutional layer optically using a new physical realization of the encoded weights, while the surrounding deterministic layers are evaluated digitally. Each pass produces one predictive sample in the form of a class-probability vector. Averaging these vectors gives the posterior-predictive distribution. We calculate the Shannon Entropy H

$$\mathrm{H} = -\sum_{c=1}^{C}\left[\frac{1}{N}\sum_{n=1}^{N} p(y_n = c \mid x, \theta_n)\cdot \log\left(\frac{1}{N}\sum_{n=1}^{N} p(y_n = c \mid x, \theta_n)\right)\right] \tag{1}$$

from this averaged distribution and the softmax entropy SE

$$\mathrm{SE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} p(y_n = c \mid x, \theta_n)\cdot \log\big(p(y_n = c \mid x, \theta_n)\big) \tag{2}$$

as the mean entropy within the individual predictive samples. Their difference, MI=H-SE, is the mutual information. The softmax entropy captures uncertainty within individual predictions and therefore serves as a measure of aleatoric uncertainty, whereas the mutual information captures disagreement among predictions generated by different weight realizations and serves as a measure of epistemic uncertainty.

We first apply the hybrid photonic BNN to blood-cell classification. The training data contain seven classes: basophils, eosinophils, immature granulocytes, lymphocytes, monocytes, neutrophils and platelets, shown in Fig. 4a [24]. We withhold erythroblasts from training and introduce them as an out-of-distribution (OOD) class during prediction. SVI infers the probabilistic weights exclusively from the seven training classes, implementation details are provided in the Supplementary Information [17]. The posterior standard deviations of representative weights evolve during training and converge to distinct values, illustrating how the network learns the stochasticity of the probabilistic convolutional layer (Fig. 4b). The experimental predictions reveal different behaviour for in-distribution and withheld images. For a representative platelet image, the network produces consistent class probabilities across all ten stochastic forward passes and therefore negligible mutual information (Fig. 4e). For a withheld erythroblast image, the predicted class changes across weight realizations, producing high mutual information and indicating epistemic uncertainty (Fig. 4f). We therefore use mutual information to detect withheld erythroblast images. Varying the rejection threshold yields an area under the receiver operating characteristic curve (AUROC) of 91.16 % shown in Fig. 4c. At a mutual-information threshold of 0.0185, rejecting uncertain images increases the classification

accuracy among retained in-distribution images from 90.26 % to 94.62 %. Figure 4d shows the corresponding confusion matrix for in-distribution and OOD images and the predicted labels or rejections.

Apart from detecting uncertainty in safety critical tasks, the hybrid photonic BNN further enables uncertainty reasoning, distinguishing epistemic novelty from aleatoric ambiguity. We train the network exclusively on MNIST and introduce two additional datasets only during inference, shown in Fig. 5a. Fashion-MNIST provides semantically unfamiliar OOD images associated with epistemic uncertainty, whereas Ambiguous-MNIST provides images that are intrinsically compatible with more than one digit class and therefore contain aleatoric uncertainty[15]. Unlike protocols that include ambiguous images during training [15, 25, 26], this procedure exposes the network to neither ambiguous nor OOD examples while it learns the weight distributions, faithfully mimicking real-world settings. Across ten predictive samples, a representative MNIST image produces consistent and confident class probabilities (Fig. 5b). A Fashion-MNIST image instead produces disagreement among predictive samples and consequently high mutual information (Fig. 5c). For an Ambiguous-MNIST image, the class-probability distribution remains broad within the individual predictive samples, producing high softmax entropy (Fig. 5d). The three datasets consequently occupy largely distinct regions of mutual-information–softmax-entropy space, separating epistemic novelty from aleatoric ambiguity (Fig. 5e). The experimentally hybrid photonic BNN achieves an MNIST classification accuracy of 96.01 %. Mutual-information-based rejection at a threshold of 0.003 08 increases the accuracy among non-rejected MNIST images to 99.70 %. Epistemic-uncertainty detection using Fashion-MNIST achieves an AUROC of 84.42 %, whereas aleatoric-uncertainty detection using Ambiguous-MNIST achieves an AUROC of 88.03 % (Fig. 5f).

Together, these experiments show that physically sampled probabilistic convolutions support both uncertainty-aware rejection and the separation of epistemic and aleatoric uncertainty in a hybrid photonic Bayesian neural network.

Discussion

Spectral encoding turns broadband chaotic light into a programmable representation of probabilistic weights and thereby extends photonic uncertainty processing from stochastic elements to a complete probabilistic convolution accelerator. Previous studies proposed photonic Bayesian neural networks based on programmed optical noise, demonstrated stochastic photonic neurons or binary probabilistic states, and used experimentally characterized photonic devices to simulate Bayesian-network models [13, 27, 28, 14, 12]. Here, the optical power and bandwidth of each spectral channel program the mean and standard deviation of a weight distribution, while the same channel continuously generates new physical realizations and participates in the convolution. A PyTorch-compatible interface embeds this operation directly in an AI workflow and replaces the software operator with the physical

accelerator during experimental prediction. To our knowledge, this combination provides the first experimental photonic implementation that unifies programmable probabilistic weights, optical convolution and end-to-end uncertainty-aware inference. The accelerator produces one probabilistic convolution output every 37.5 ps. Although rate definitions differ between architectures, this operating interval is substantially shorter than the nanosecond-scale stochastic operation reported or projected for other photonic probabilistic systems [27, 28]. Spectral encoding therefore advances the state of the art not only by accelerating random sampling, but by storing, sampling and processing probabilistic weights within the same optical signal.

The system-level value of the accelerator follows from the observation that Bayesian neural networks do not require every model parameter to be stochastic. Partially stochastic networks can retain expressive predictive distributions and match or outperform fully stochastic models while reducing their memory and computational requirements [29]. We exploit this principle through hardware–algorithm co-design. An architecture search concentrates stochasticity in one probabilistic convolutional layer and favours grouped and depthwise-separable operations that maximize predictive performance for a given number of photonic kernels. The nine simultaneously encoded weights form a complete 3×3 probabilistic kernel, while convolutional weight sharing reuses each programmed distribution across many spatial positions. The physical resource requirement therefore depends on the number of concurrently evaluated weights rather than on the total number of convolution operations. A differentiable abstraction of the accelerator further enables digital training without repeated access to the physical system. This approach follows the broader principle of hardware-aware training, in which calibrated device behaviour and non-idealities are represented in software and included during optimization [30]. During experimental prediction, we replace the abstract operator with the photonic accelerator and observe close agreement between their predictive performance. The resulting hybrid network detects withheld blood-cell classes and improves the accuracy of retained in-distribution images, while evaluation on the Dirty-MNIST uncertainty benchmark separates epistemic novelty from aleatoric ambiguity [24, 15]. The selective stochasticity, weight reuse and hardware-aware training make the photonic probabilistic convolution accelerator a practical system component for uncertainty-aware AI workloads.

The photonic probabilistic convolution accelerator offers a practical route towards full integration and larger-scale operation. The integrated chirped waveguide Bragg grating already replaces kilometre-scale fibre delays with a compact on-chip dispersive element, substantially reducing optical propagation time and latency compared with earlier frequency–time-interleaved processors [11]. The remaining optical functions can likewise be integrated. Erbium-doped silicon-nitride waveguides or electrically pumped gain media can provide broadband entropy sources [31, 32], while integrated electro-optic modulators and photodetectors operate at bandwidths exceeding 100 GHz [33, 34]. Dense wavelength-parallel processors and integrated spectral shapers further provide a route to programming more frequency channels with lower reconfiguration overhead [35, 36, 37]. Combining these components in an end-to-end integrated system would further reduce footprint and latency.

Our results demonstrate that spectrally encoded probabilistic weights transform intrinsic optical noise into a programmable computational resource, enabling high-speed uncertainty reasoning and establishing a path towards uncertainty-aware photonic artificial intelligence.

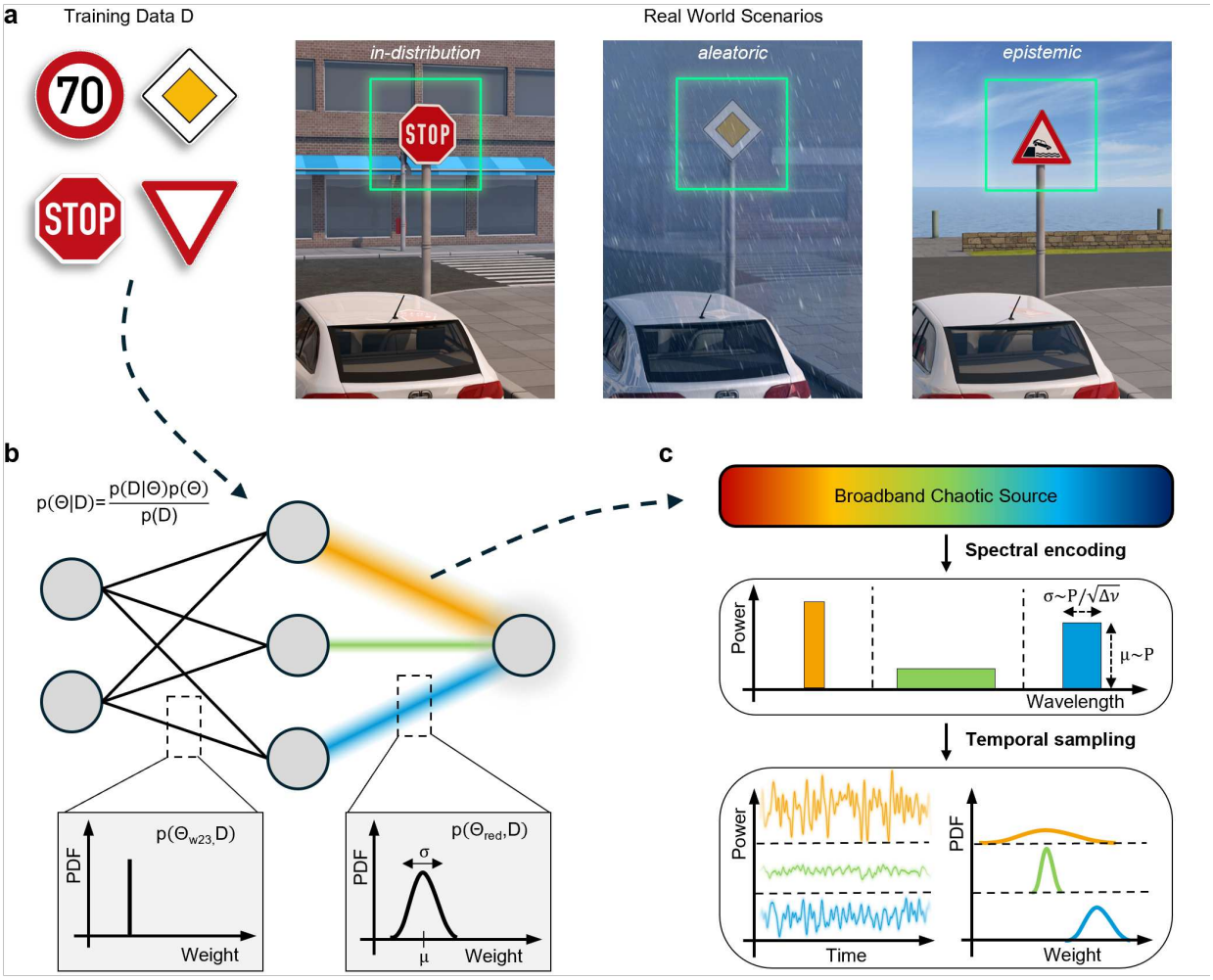


Fig. 1: Uncertainty-aware photonic artificial intelligence. **a**, In real-world applications, AI systems encounter aleatoric uncertainty, for example due to limited vision, and epistemic uncertainty arising from completely unknown inputs. Quantifying these uncertainties is crucial for deploying AI in safety-critical environments. **b**, Bayesian neural networks infer probability distributions over model parameters from the training data, whereas conventional networks infer single point estimates. This probabilistic formulation provides a principled framework for uncertainty reasoning. Only a subset of network weights needs to be stochastic [29]. Inference requires sampling from these distributions and repeating the forward pass multiple times, which incurs substantial computational overhead. **c**, We encode the weight distributions in the emission spectrum of a broadband chaotic light source and map each distribution to an individual frequency channel. Random beating between the frequency components within each channel produces temporal power fluctuations. We program the mean of the fluctuations by setting the total power P in the frequency channel and program the standard deviation by setting the optical bandwidth $\Delta\nu$ of the channel. In this way, a single source both stores and samples from multiple weight distributions in parallel.

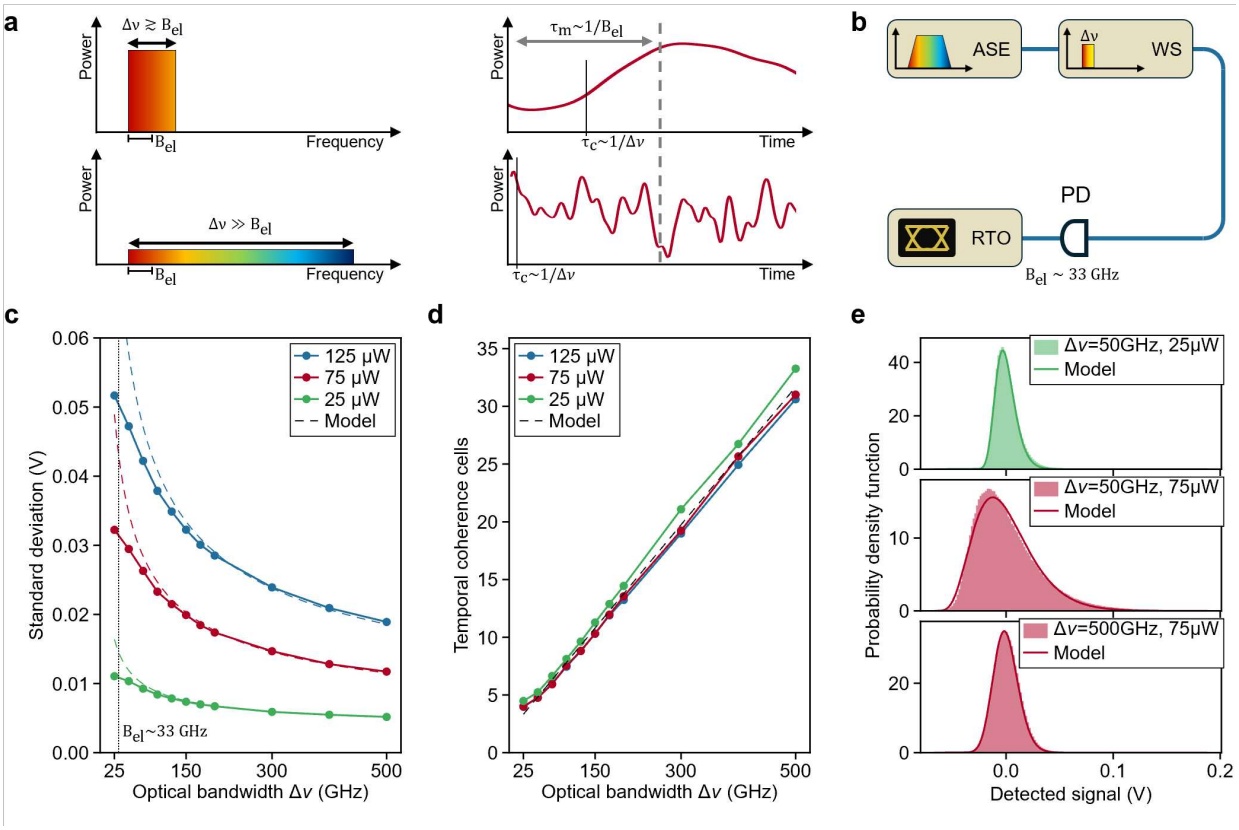


Fig. 2: Spectral encoding of probability distributions. **a**, Random beating between the frequency components of chaotic light produces temporal power fluctuations with a coherence time $\tau_c \sim 1/\Delta\nu$, where $\Delta\nu$ is the optical bandwidth. The finite electrical bandwidth B_{el} of the electronic detection and readout system defines an effective measurement interval $\tau_m \sim 1/B_{el}$. For a narrow optical channel, the detector resolves the slowly varying fluctuations. Increasing the optical bandwidth shortens τ_c, so that each measurement averages over multiple effective temporal coherence cells M_{eff} and suppresses the detected fluctuations. **b**, We generate chaotic light through amplified spontaneous emission (ASE) in an erbium-doped fibre, program its spectrum with a waveshaper (WS), and record the power fluctuations using a 38 GHz photodetector (PD) and an 80 GSa s^{-1} real-time oscilloscope (RTO) with a bandwidth of 33 GHz. **c**, We record RTO voltage traces at different optical powers and bandwidths. Increasing the optical bandwidth reduces the resolved fluctuations because the detector averages over more temporal coherence cells. At large optical bandwidths, the standard deviation follows $1/\sqrt{\Delta\nu}$ and eventually approaches the electronic-noise floor. Increasing the optical power increases the optical fluctuations linearly. **d**, We extract M_{eff} from the measured voltage distributions at each operating point. M_{eff} increases approximately linearly with optical bandwidth and remains largely independent of optical power, identifying bandwidth as the principal control parameter for temporal averaging. **e**, Representative measured probability densities. At a fixed bandwidth of 50 GHz, increasing the optical power from 25 μW to 75 μW proportionally increases the standard deviation from 0.010 V to 0.029 V. At a fixed power of 75 μW, increasing the bandwidth from 50 GHz to 500 GHz narrows the distribution to a standard deviation of 0.012 V and makes it more Gaussian-like.

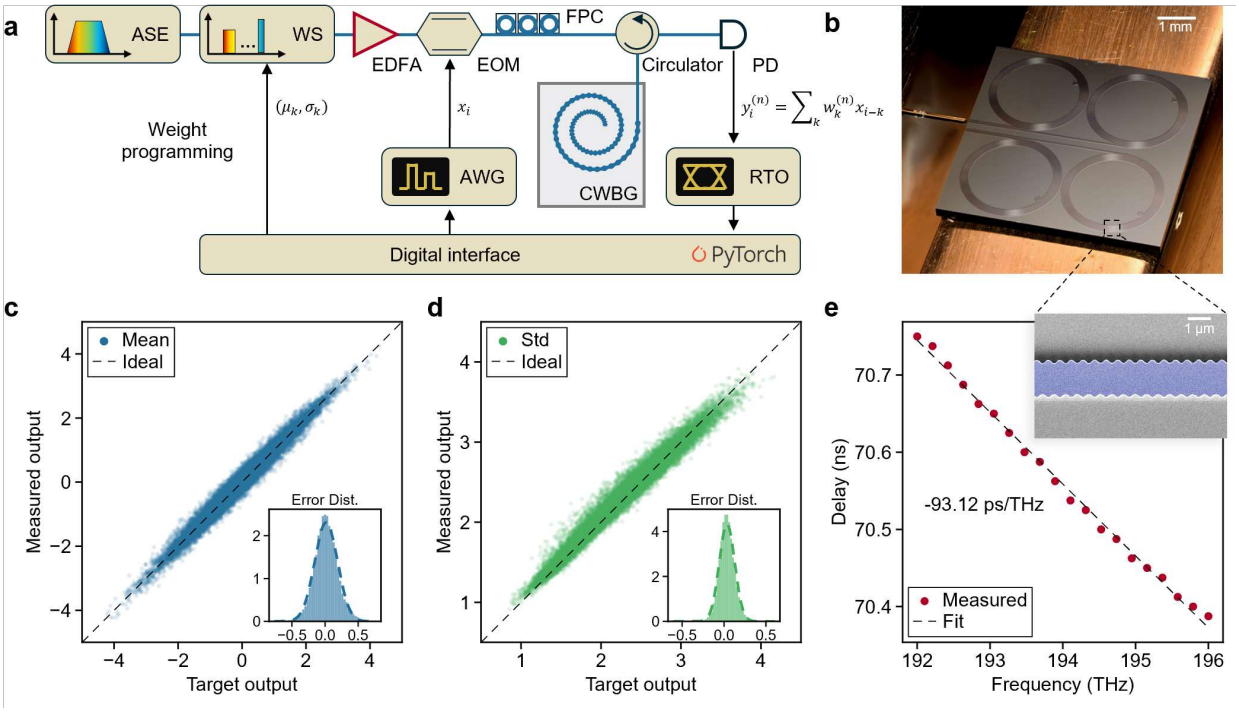


Fig. 3: High-speed photonic convolution processing with probabilistic weights. **a**, We combine spectral weight encoding with frequency–time interleaving to realize a high-speed photonic probabilistic convolution accelerator. A waveshaper (WS) programs nine probabilistic weights in frequency channels separated by 403 GHz within the emission spectrum of an amplified spontaneous emission (ASE) source. An 8-bit arbitrary-waveform generator (AWG) operating at 80 GSa s^{-1} drives an electro-optic modulator (EOM), which encodes the temporal input data simultaneously onto all channels. Each input value is represented by a three-sample symbol with a duration of 37.5 ps. A chirped waveguide Bragg grating (CWBG) delays neighbouring channels by one symbol and aligns their contributions at the photodetector to implement a probabilistic convolution. A fibre polarization controller (FPC) sets the input polarization to the transverse-electric mode. An 8-bit real-time oscilloscope (RTO) operating at 80 GSa s^{-1} records the output, while a PyTorch-compatible digital interface controls weight programming, waveform generation and data acquisition. **b**, Photograph of the integrated CWBG on a low-loss silicon-nitride platform. The on-chip dispersive element provides the frequency-dependent delay required for convolution while maintaining a short optical path. **c,d**, We characterize the accelerator using 25 randomly generated probabilistic kernels. For each kernel, we evaluate a random input vector of length 1000 and acquire 1500 samples from each output distribution. The measured output means (**c**) and standard deviations (**d**) closely follow their target values. The errors, defined as measured minus target, have mean ± s.d. values of 0.016 ± 0.173 for the output mean and 0.036 ± 0.092 for the output standard deviation. **e**, We measure the delay between the AWG output and the RTO input as a function of optical frequency. A linear fit yields a frequency-dependent delay of −93.12 ps THz^{-1}, corresponding to a delay of 37.5 ps between adjacent frequency channels. Inset, scanning electron micrograph of the chirped grating; its period varies along the propagation direction so that the centre of the reflection band shifts along the grating.

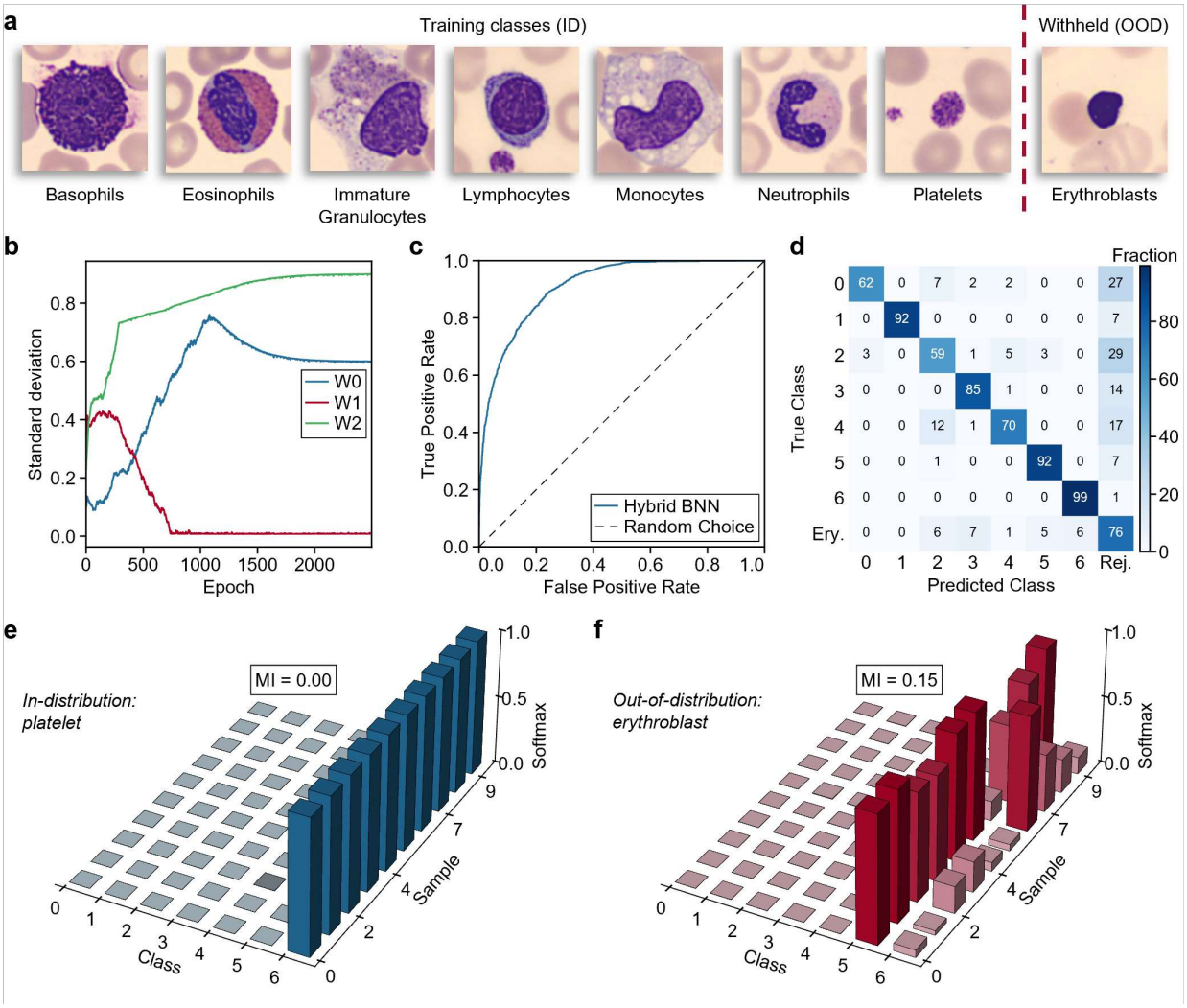


Fig. 4: Uncertainty-aware blood-cell classification. **a**, Representative microscope images from the seven blood-cell classes used for training. We withhold erythroblasts from training and introduce them as an out-of-distribution (OOD) class during inference. **b**, Evolution of the posterior standard deviations of three representative weights in the probabilistic convolutional layer during stochastic variational inference. **c**, Receiver operating characteristic (ROC) curve for detecting withheld erythroblast images using predictive mutual information (MI), yielding an area under the ROC curve (AUROC) of 91.16 %. The dashed line denotes random choice. **d**, Row-normalized confusion matrix at an MI rejection threshold of 0.0185. The final row denotes withheld erythroblast images, and the final column denotes rejected samples. At this MI threshold, the classification accuracy among retained in-distribution images increases from 90.26 % to 94.62 %. **e,f**, Class probabilities from ten stochastic forward passes of the hybrid photonic BNN for a representative in-distribution platelet image (**e**) and a withheld erythroblast image (**f**). In each pass, we evaluate the probabilistic convolutional layer optically using a new physical realization of the encoded weights and evaluate the surrounding deterministic layers digitally. The in-distribution image yields consistent predictions and negligible MI, whereas disagreement among predictions for the erythroblast produces high MI.

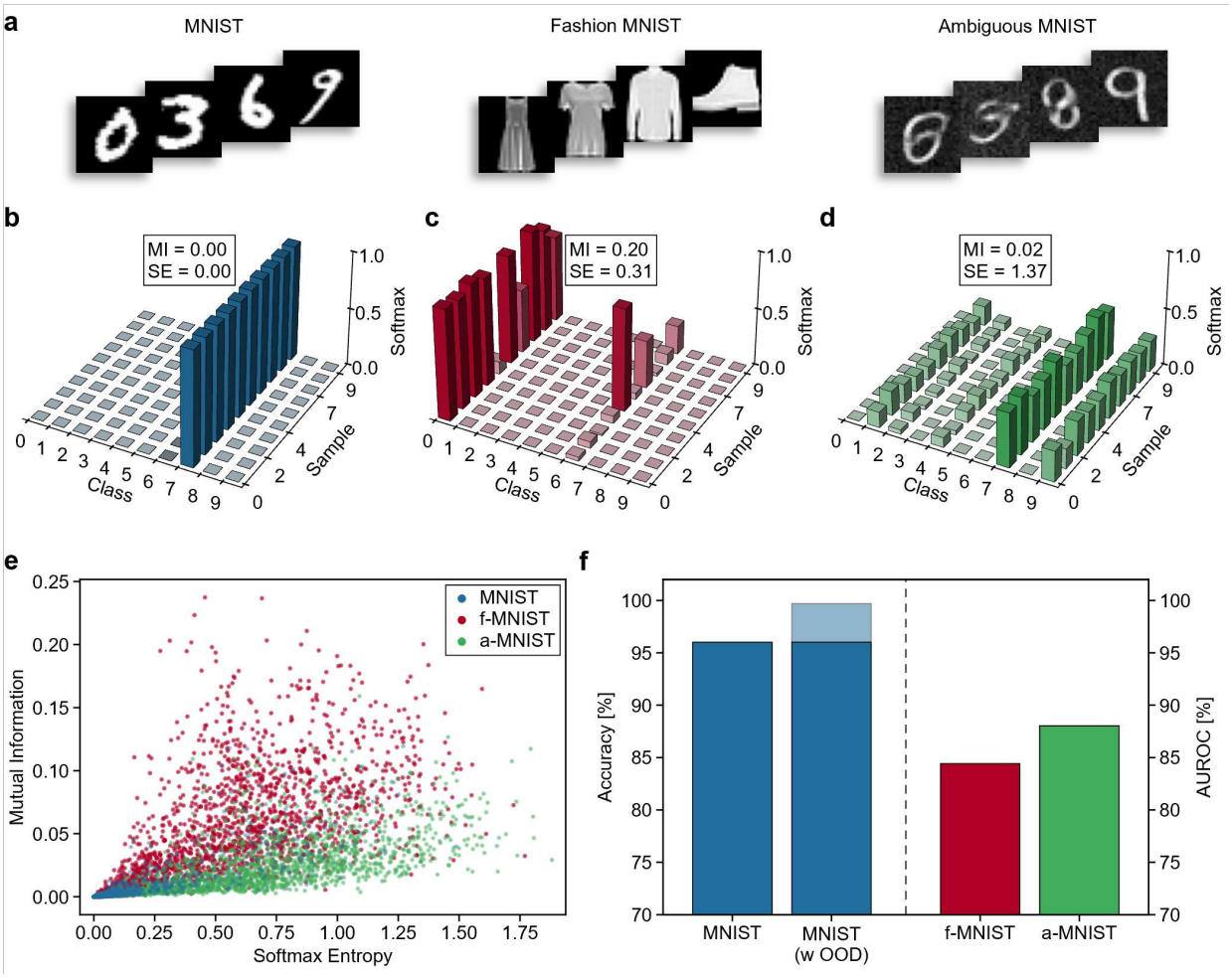


Fig. 5: Disentanglement of aleatoric and epistemic uncertainty. **a**, Representative images from MNIST, Fashion-MNIST and Ambiguous-MNIST. We train the hybrid photonic BNN exclusively on MNIST. Fashion-MNIST provides out-of-distribution images associated with epistemic uncertainty, whereas Ambiguous-MNIST provides intrinsically ambiguous images associated with aleatoric uncertainty. **b–d**, Class probabilities from ten predictive samples of the hybrid photonic BNN for representative MNIST (**b**), Fashion-MNIST (**c**) and Ambiguous-MNIST (**d**) images. Each sample corresponds to one stochastic forward pass with a new physical realization of the probabilistic weights. We perform all probabilistic convolutions optically and compute all other deterministic operations digitally. The in-distribution image yields consistent, confident predictions across samples. The out-of-distribution image produces disagreement among samples and therefore high mutual information (MI); the ambiguous image produces high entropy within individual samples, reflected in high softmax entropy (SE). **e**, Mutual information versus softmax entropy for individual test images. The three datasets occupy distinct regions of uncertainty space, enabling the separation of epistemic novelty from aleatoric ambiguity. **f**, Experimental task performance of the hybrid photonic BNN. Mutual-information-based rejection at a threshold of 0.003 08 increases the MNIST classification accuracy from 96.01 % to 99.70 % for non-rejected test images. Epistemic- and aleatoric-uncertainty detection achieve AUROCs of 84.42 % for Fashion-MNIST and 88.03 % for Ambiguous-MNIST, respectively.

Materials and Methods

Experimental Setup

We house the photonic accelerator and its optical and electronic subsystems in a standard 19-inch rack. An Agiltron ASES-1611A3113 amplified-spontaneous-emission source generates broadband chaotic light. A Coherent WaveShaper 1000B programs the probabilistic weights by setting the power and bandwidth of the individual frequency channels. We subsequently amplify the shaped spectrum using the preamplifier stage of a PriTel LNTFA-20-NMA-DP erbium-doped fibre amplifier, which increases the optical power with low excess noise. An Optilab IML-1550-40-PM-V electro-optic modulator encodes the analogue electronic input waveform onto the shaped optical spectrum.

The modulated signal passes through a fibre polarization controller and an optical circulator before entering the integrated chirped waveguide Bragg grating (CWBG) shown in Fig. 3). The polarization controller sets the input to the transverse-electric mode of the CWBG, which introduces a frequency-dependent delay of 37.5 ps between adjacent frequency channels. The circulator separates the reflected convolution signal from the optical input. A Thorlabs RXM38AF photoreceiver detects the summed optical output. The photoreceiver comprises a high-speed photodiode with a responsivity of $0.7\,\mathrm{A\,W^{-1}}$ and a tunable-gain transimpedance amplifier operated at a total conversion gain of $1000\,\mathrm{V\,W^{-1}}$.

A Keysight M8196A arbitrary-waveform generator (AWG) provides the digital-to-analogue interface, and a Keysight DSAX95004Q real-time oscilloscope (RTO) provides the analogue-to-digital interface. Both instruments operate at $80\,\mathrm{GSa\,s^{-1}}$. We use pulse-amplitude modulation with three samples per symbol, corresponding to a symbol duration of 37.5 ps. A Xilinx ZCU216 board generates phase-locked 100 MHz and 2.5 GHz reference clocks that synchronize the AWG and RTO and align their sampling rates. Python scripts control the waveshaper, AWG and RTO and expose the photonic accelerator through a PyTorch-compatible interface.

Chirped Grating Fabrication

We fabricate the CWBGs on a wafer-scale, 200 nm-thick Si_3N_4 platform [38]. Fabrication starts from a 4-inch silicon wafer with a 10 µm wet-oxide SiO_2 layer that forms the lower cladding. We deposit a 200 nm Si_3N_4 layer by low-pressure chemical-vapour deposition and anneal it at 1200 °C to reduce hydrogen-related optical loss. We define the waveguides by deep-ultraviolet lithography (KrF 248 nm) and transfer the pattern by anisotropic reactive-ion etching with fluorine-based gases. We then encapsulate the waveguides in a 3 µm hydrogen-free, low-loss SiO_2 upper cladding [39]. We separate the dies by deep reactive-ion etching through the oxide and partially into the silicon substrate, followed by backside grinding.

The CWBG contains 113 000 grating periods that vary linearly from 496.7 nm to 509.3 nm. The

resulting 56.8 mm-long device is folded into a spiral to reduce its footprint. We choose a sufficiently large spiral radius to limit radius-dependent variations in the effective refractive index [40]. The waveguide has an average width of 1.6 µm, close to single-mode operation, and a periodic width corrugation that provides distributed Bragg reflection. A photograph of the integrated CWBG is shown in Fig. 3b, and the inset of Fig. 3e shows a scanning electron micrograph of the corrugated waveguide.

We suppress reflections from the grating boundaries by apodizing the corrugation amplitude with a hyperbolic-tangent profile [40, 41]. Sharp waveguide bends terminate the distal end of the CWBG and further suppress end-facet reflections. We simulate the optical response using the transfer-matrix method [42], representing each grating period by a 2×2 transfer matrix. To reduce the computational cost, we divide the grating into groups of 100 consecutive periods and assume identical matrices within each group. Figure 3e shows the measured frequency-dependent delay of the fabricated device. We package the CWBG with an ultra-high-numerical-aperture silica fibre (UHNA7) for integration into the rack-mounted photonic setup.

Neural Network Training

We implement and train the Bayesian neural networks in PyTorch [20, 21]. Pyro provides stochastic variational inference (SVI) [23], and Brevitas provides quantization-aware training [43]. We combine these frameworks in a custom, SVI-compatible probabilistic convolution operator that represents the photonic accelerator during training. The operator models the physical convolution statistics together with relevant hardware constraints, including digital-to-analogue converter quantization, the analogue-to-digital converter range and stochastic deviations in weight programming. We use the tensor-manipulation library Einops to implement the custom depthwise-separable (DWS) convolutions and associated tensor rearrangements [44].

We train the networks on a heterogeneous compute cluster managed with SLURM [45]. The available CPUs comprise AMD EPYC 7452, AMD EPYC 7302P and Intel Xeon E5-2667 v3 processors. The corresponding GPU resources comprise NVIDIA A30, NVIDIA GeForce RTX 2080 Ti and NVIDIA Tesla K80 devices. We manage and document the training runs using the open-source SEML framework, which stores configurations and results in an on-premises MongoDB database. A custom SEML extension implements the asynchronous successive-halving algorithm (ASHA) as an early-stopping procedure for efficient hyperparameter optimization [46].

During experimental inference, we map the two-dimensional probabilistic convolutions used by the network onto the hardware-compatible one-dimensional convolutions evaluated by the photonic accelerator. Custom data-reordering functions transform the input tensors and programmed kernels before optical execution and restore the two-dimensional output layout afterwards. We compile these functions just in time using Numba [47]. This implementation maximizes data reuse while maintaining a common PyTorch interface for software training and physical prediction.

References and Notes

1. Amodei, D. *et al.* Concrete problems in AI safety (2016). ArXiv:1606.06565.

2. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, vol. 30, 6402–6413 (2017).

3. Kendall, A. & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, vol. 30, 5574–5584 (2017).

4. Lambert, B., Forbes, F., Doyle, S., Dehaene, H. & Dojat, M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine* **150**, 102830 (2024).

5. Jospin, L. V., Laga, H., Boussaid, F., Buntine, W. & Bennamoun, M. Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine* **17**, 29–48 (2022).

6. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *Proceedings of Machine Learning Research*, 1613–1622 (2015).

7. Pedrajas, S. P., Resano, J. & Gracia, D. S. Bnnrv: Hardware and software optimizations for weight sampling in bayesian neural networks on edge risc-v cores. *IEEE Transactions on Circuits and Systems for Artificial Intelligence* **3**, 78–89 (2026).

8. Liu, S. *et al.* Bayesian neural networks using magnetic tunnel junction-based probabilistic in-memory computing. *Frontiers in Nanotechnology* **4**, 1021943 (2022).

9. Lin, Y. *et al.* Uncertainty quantification via a memristor Bayesian deep neural network for risk-sensitive reinforcement learning. *Nature Machine Intelligence* **5**, 714–723 (2023).

10. Le Gallo, M. *et al.* A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics* **6**, 680–693 (2023).

11. Xu, X. *et al.* 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).

12. Brückerhoff-Plückelmann, F. *et al.* Probabilistic photonic computing with chaotic light. *Nature Communications* **15**, 10445 (2024).

13. Wu, C., Yang, X., Chen, Y. & Li, M. Photonic bayesian neural network using programmed optical noises. *IEEE Journal of Selected Topics in Quantum Electronics* **29**, 1–6 (2023).

14. Zhuge, Y. *et al.* Photonic bayesian neural networks: Leveraging programmable noise for robust and uncertainty-aware computing. *Advanced Science* **12**, 2500525 (2025).

15. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H. S. & Gal, Y. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24384–24394 (2023).

16. Goodman, J. W. Statistical optics. *New York, Wiley-Interscience, 1985, 567 p.* **1** (1985).

17. Materials and methods are available as supplementary material.

18. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708 (2017).

19. Howard, A. G. *et al.* MobileNets: Efficient convolutional neural networks for mobile vision applications (2017). ArXiv:1704.04861.

20. Paszke, A. *et al.* Automatic differentiation in PyTorch. In *Proceedings of the NIPS 2017 Workshop on Autodiff* (2017).

21. Ansel, J. *et al.* PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, 929–947 (2024).

22. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research* **14**, 1303–1347 (2013).

23. Bingham, E. *et al.* Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research* **20**, 1–6 (2019).

24. Yang, J. *et al.* MedMNIST v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data* **10**, 41 (2023).

25. Sun, H., van Breugel, B., Crabbé, J., Seedat, N. & van der Schaar, M. What is flagged in uncertainty quantification? latent density models for uncertainty categorization. In *Advances in Neural Information Processing Systems*, vol. 36, 4664–4684 (2023).

26. Steger, S., Knoll, C., Klein, B., Fröning, H. & Pernkopf, F. Function space diversity for uncertainty prediction via repulsive last-layer ensembles. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling* (2024).

27. Ma, B., Zhang, J., Li, X. & Zou, W. Stochastic photonic spiking neuron for bayesian inference with unsupervised learning. *Optics Letters* **48**, 1411–1414 (2023).

28. Choi, S. *et al.* Photonic probabilistic machine learning using quantum vacuum noise. *Nature Communications* **15**, 7760 (2024).

29. Sharma, M., Farquhar, S., Nalisnick, E. & Rainforth, T. Do Bayesian neural networks need to be fully stochastic? In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, vol. 206 of *Proceedings of Machine Learning Research*, 7694–7722 (2023).

30. Rasch, M. J. *et al.* Hardware-aware training for large-scale and diverse deep learning inference workloads using in-memory computing-based accelerators. *Nature Communications* **14**, 5282 (2023).

31. Liu, Y. *et al.* A photonic integrated circuit-based erbium-doped amplifier. *Science* **376**, 1309–1313 (2022).

32. Shi, B., Calabretta, N. & Stabile, R. InP photonic integrated multi-layer neural networks: Architecture and performance analysis. *APL Photonics* **7**, 010801 (2022).

33. Weigel, P. O. *et al.* Bonded thin-film lithium niobate modulator on a silicon photonics platform exceeding 100 GHz bandwidth. *Optics Express* **26**, 23728–23739 (2018).

34. Shi, Y., Li, X., Zou, M., Yu, Y. & Zhang, X. 103 GHz germanium-on-silicon photodiode enabled by an optimized U-shaped electrode. *Photonics Research* **12**, 1–6 (2024).

35. Metcalf, A. J. *et al.* Integrated line-by-line optical pulse shaper for high-fidelity and rapidly reconfigurable RF filtering. *Optics Express* **24**, 23925–23940 (2016).

36. Shu, H., Chang, L., Tao, Y. *et al.* Microcomb-driven silicon photonic systems. *Nature* **605**, 457–463 (2022).

37. Bai, B., Yang, Q., Shu, H. *et al.* Microcomb-based integrated photonic processing unit. *Nature Communications* **14**, 66 (2023).

38. Ji, X. *et al.* Efficient mass manufacturing of high-density, ultra-low-loss Si_3N_4 photonic integrated circuits. *Optica* **11**, 1397–1407 (2024).

39. Qiu, Z. *et al.* Hydrogen-free low-temperature silica for next generation integrated photonics (2023). ArXiv:2312.07203.

40. Du, Z. *et al.* Silicon nitride chirped spiral bragg grating with large group delay. *APL Photonics* **5**, 101302 (2020).

41. Li, Y. *et al.* Large group delay and low loss optical delay line based on chirped waveguide bragg gratings. *Optics Express* **31**, 4630–4638 (2023).

42. Erdogan, T. Fiber grating spectra. *Journal of Lightwave Technology* **15**, 1277–1294 (1997).

43. Franco, G., Pappalardo, A. & Fraser, N. J. Xilinx/brevitas. Zenodo (2025).

44. Rogozhnikov, A. Einops: Clear and reliable tensor manipulations with einstein-like notation. In *International Conference on Learning Representations* (2022).

45. Klusáček, D., Corbalán, J. & Rodrigo, G. P. (eds.) *Job Scheduling Strategies for Parallel Processing*, vol. 14283 of *Lecture Notes in Computer Science* (Springer, 2023).

46. Li, L. *et al.* A system for massively parallel hyperparameter tuning. In Dhillon, I., Papailiopoulos, D. & Sze, V. (eds.) *Proceedings of Machine Learning and Systems*, vol. 2, 230–246 (2020).

47. Lam, S. K., Pitrou, A. & Seibert, S. Numba: A LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, 1–6 (2015).

48. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877 (2017).

49. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233 (1999).

50. Roth, W. *et al.* Resource-efficient neural networks for embedded systems. *Journal of Machine Learning Research* **25**, 1–51 (2024).

51. Wang, X., Borras, H., Klein, B. & Fröning, H. On hardening DNNs against noisy computations. In *Proceedings of the 7th Workshop on Machine Learning and Programming Languages (AccML), co-located with HiPEAC* (2025). ArXiv:2501.14531.

52. Tempczyk, P., Smoczyński, K., Smolenski-Jensen, P. & Cygan, M. One simple trick to fix your Bayesian neural network (2022). ArXiv:2207.13167.

53. Müller, S. G. & Hutter, F. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 774–782 (2021).

54. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H. S. & Gal, Y. Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty (2021). ArXiv:2102.11582.

55. Defazio, A. *et al.* The road less scheduled. In *Advances in Neural Information Processing Systems*, vol. 37, 9974–10007 (2024).

56. Liu, L. *et al.* On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)* (2020).

Acknowledgments

We thank Jochen Stuhrmann, from Illustrato, for his assistance with the illustrations. The samples were partially fabricated in the EPFL Center of MicroNanoTechnology (CMi).

Funding: This work was supported by funding from the European Research Council grant 101200429 (WP) and Contract N660012424006 (NaPSAC) from the Defense Advanced Research Projects Agency (DARPA), Defense Sciences Office (DSO) (TK, WP).

Author contributions: Conceptualization: FBP, HB, LM, BK
Methodology: FBP, JH, SH, XJ, FE, JD, LM, PS, JS
Investigation: FBP, HB, LM, JH
Visualization: FBP, HB, SH, XJ
Funding acquisition: TK, HF, WP
Project administration: FBP, TK, HF, WP
Supervision: TK, HF, WP
Writing – original draft: FBP, HB
Writing – review & editing: All authors

Competing interests: There are no competing interests to declare.

Data and materials availability: All data are available in the main text or the supplementary materials.

Supplementary materials

Supplementary Text
Figs. S1 to S6
Table S1

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BNNSupplementaryInformation.pdf