Research Article Vol. 16, No. 11 / 1 Nov 2025 / Biomedical Optics Express 4312
Biomedical Optics EXPRESS

MediONN: an integrated photonic chip optical neural network for deep learning-based semantic segmentation in early detection of pancreatic cancer

CHUN-JU YANG,[1] HANQING ZHU,[1] SHUPENG NING,[1] JIAQI GU,[1] CHENGHAO FENG,[1] DAVID Z. PAN,[1] AND RAY T. CHEN[1,2,*]

[1]*Microelectronics Research Center, Electrical and Computer Engineering Department, University of Texas at Austin, Austin, TX 78758, USA*
[2]*Omega Optics Inc., Austin, TX 78759, USA*
[*]*raychen@uts.cc.utexas.edu*

Abstract: Pancreatic cancer remains one of the deadliest cancers due to the lack of effective early detection tools. While deep neural networks (DNNs) have shown promise in tumor segmentation, electronic accelerators suffer from power inefficiency and latency. To address this, we propose MediONN—a photonic neural network system implemented on an integrated chip, optimized for 3D medical image segmentation. MediONN integrates a 4×4 photonic neural processor within a hierarchical 3D optical computation framework. To improve training convergence, we introduce a segmentation-specific Gaussian weight initialization strategy, along with 3D optical convolutional layers for volumetric feature extraction. Unlike prior photonic systems focused on classification, MediONN is the first to demonstrate optical neural networks (ONNs) directly applied to 3D segmentation. On the NIH pancreas CT dataset, MediONN achieves a Dice Similarity Coefficient (DSC) of 0.5215 (2D) and 0.5302 (3D), with peak DSCs of 0.5919 (2D) and 0.8788 (3D). Comprehensive evaluation metrics confirm MediONN's segmentation accuracy is comparable to electronic counterparts, while offering significant gains in computational speed and energy efficiency. These results highlight the scalability and biomedical potential of integrated photonic ONNs.

© 2025 Optica Publishing Group under the terms of the Optica Open Access Publishing Agreement

1. Introduction

Pancreatic cancer has the lowest five-year survival rate among all cancers and is projected to become the second leading cause of cancer-related deaths in the United States by 2030 [1]. Despite advances in imaging, progress in improving pancreatic cancer–specific outcomes has been limited in recent decades. Because prognosis declines sharply once tumors exceed 2 cm in size, early detection is critical. Although computed tomography (CT) remains the primary imaging modality for pancreatic cancer [2], its sensitivity for small tumors is suboptimal; approximately 40% of tumors smaller than 2 cm are missed [3]. Moreover, CT interpretation is subject to inter-observer variability, depending on the experience and availability of radiologists.

Recent developments in deep learning (DL) have enabled significant progress in medical image analysis, especially in automated tumor detection and segmentation. Deep neural networks (DNNs) have demonstrated state-of-the-art performance in tasks ranging from image classification [4,5] and object detection [6] to medical diagnosis [7,8]. However, the increasing size of DNN models and the corresponding computational demand pose a challenge for traditional electronic hardware accelerators. Devices such as GPUs, FPGAs, and digital ASICs face growing limitations in power efficiency, latency, and scalability [9–11].

To address these bottlenecks, we explore Optical Neural Networks (ONNs), which leverage the parallelism and bandwidth of light to achieve low-latency, energy-efficient computation

#574635 https://doi.org/10.1364/BOE.574635
Journal © 2025 Received 25 Jul 2025; revised 18 Sep 2025; accepted 18 Sep 2025; published 6 Oct 2025

Biomedical Optics EXPRESS

[12,13]. Compared to traditional electronic accelerators such as GPUs and FPGAs, ONNs offer orders-of-magnitude improvements in throughput and energy efficiency by offloading computation into the optical domain.

However, prior ONN research has primarily focused on image classification tasks, where a single global label is predicted per image. These models typically rely on compact matrix multiplication engines and do not support dense spatial prediction. By contrast, medical image segmentation demands voxel-wise output, multi-scale spatial reasoning, and volumetric consistency—challenges that are significantly harder to solve in optical hardware due to training instability, limited spatial expressiveness, and architectural difficulties in scaling to 3D data. Existing ONN architectures are not readily extensible to segmentation, especially in the 3D biomedical domain.

To overcome these challenges, we propose *MediONN*, a photonic neural architecture specifically tailored for 3D medical image segmentation. Our system integrates a custom-designed 4×4 optical processor into a hierarchical 3D computational pipeline, combining compact photonic layers with domain-specific innovations such as segmentation-aware weight initialization and volumetric convolution. MediONN enables dense tumor region prediction in volumetric CT scans while retaining the energy and latency advantages of ONNs, establishing a new direction for photonic computing in medical applications.

In this study, we present an ONN-based segmentation model for early-stage pancreatic cancer detection, integrated into a hierarchical 3D pipeline. Using the NIH pancreas CT dataset, the proposed system achieves a DSC of 0.5215 for 2D and 0.5302 for 3D segmentation. The DSC quantifies the spatial overlap between predicted tumor regions and ground truth annotations, serving as a standard metric for evaluating segmentation accuracy. Our model also demonstrates comparable peak performance to state-of-the-art electronic systems, with significantly improved energy and speed characteristics, highlighting the promise of ONNs as a scalable solution for medical image analysis.

2. MediONN: optical neural network architecture

Our goal is to accelerate the most computation-intensive layers in neural networks—namely convolutional and fully-connected layers—by offloading them to integrated photonic hardware. Figure 1 illustrates the end-to-end training and deployment pipeline of our MediONN system, which is inspired by prior work in structured optical matrix multiplication [12,14–17]. The workflow begins in Fig. 1(a), where a Differentiable Photonic Emulator (DPE)—a neural network-based surrogate model—is trained to mimic the physical behavior of the silicon photonic chip. To construct the DPE, we first perform empirical calibration of the actual hardware, including measurements such as power transmission spectra, phase-voltage relationships of modulators, thermal crosstalk coefficients, and insertion loss profiles. These measured responses serve as supervision signals to train the DPE via regression. Once trained, the DPE provides a differentiable, hardware-aware forward function that predicts the output of the photonic computing module when executing specific matrix-vector multiplication (MVM) operations, including convolution filters and fully-connected projections. To further enhance robustness under low-precision control, we extend the DPE-based training framework with quantization-aware optimization and dynamic noise injection techniques. Specifically, during training, forward propagation simulates the limited resolution of device control (e.g., via rounding operations), while random perturbations are injected into Mach–Zehnder interferometer (MZI) phase settings to mimic quantization noise and hardware variability. This encourages the model to learn parameters that remain robust under physical imprecision.

This emulated hardware model is then integrated into the full training loop. During training, the neural network model intended for deployment (e.g., a segmentation model) is simulated with

Biomedical Optics EXPRESS

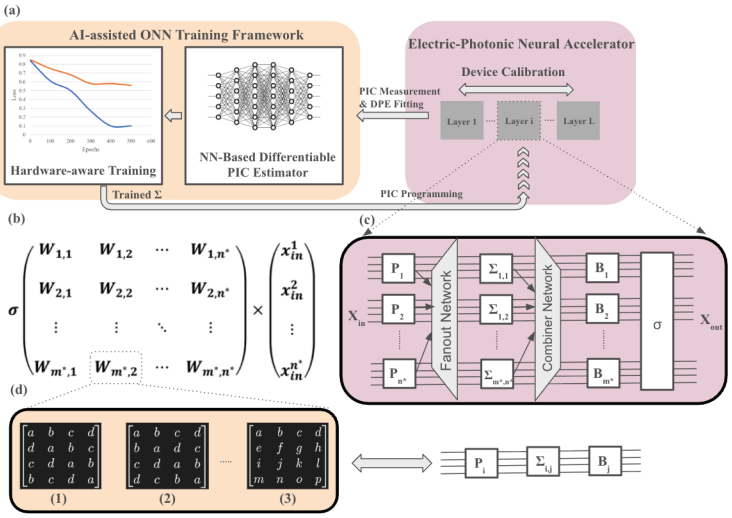


Fig. 1. Overview of the optical neural network architecture. (a) shows the training framework; (b) represents the matrix block structure; (c) shows the hardware layout for one layer with projection (P), transformation (B), and diagonal (Σ) units; and (d) visualizes alternative unitary transformations used in programmable photonic meshes.

the DPE in place of the photonic chip. In this way, model parameters in our optical blocks—are optimized in software while accounting for physical non-idealities.

Because the full weight matrix $\mathbf{W} \in \mathbb{C}^{m \times n}$ cannot be directly encoded in the chip, we partition it into a grid of smaller $k \times k$ submatrices $\mathbf{W}_{i,j}$, each of which corresponds to an independent photonic computing unit. Figure 1(b) illustrates this blockwise matrix structure and the associated MVM process, including signal fanout, phase modulation, and recombination components specifically optimized for the 4×4 case ($k = 4$). Each output segment is computed as a sum of the products of submatrices and input slices:

$$\mathbf{x}'_{\text{out}} = \mathbf{W}\mathbf{x}_{\text{in}} = \begin{pmatrix} \sum_{j=1}^{n^*} W_{1,j}\mathbf{x}_{\text{in}}^j \\ \sum_{j=1}^{n^*} W_{2,j}\mathbf{x}_{\text{in}}^j \\ \vdots \\ \sum_{j=1}^{n^*} W_{m^*,j}\mathbf{x}_{\text{in}}^j \end{pmatrix}$$

Figure 1(c) shows the optical hardware realization of one such submatrix $W_{i,j}$. Each block is implemented as $W_{i,j} = B\Sigma_{i,j}P$, where P and B are fixed unitary transformations realized using passive butterfly-style meshes, and $\Sigma_{i,j}$ is a trainable diagonal matrix implemented by MZI-based optical attenuators. Our current implementation supports 4×4 matrix blocks, which serve as the basic compute units for optical MVMs.

Compared to traditional $k \times k$ mesh-based MZI arrays, our architecture significantly reduces optical component count. The use of logarithmic-depth networks reduces complexity to $O(k \log_2 k)$ for passive components, while training is limited solely to the diagonal $\Sigma_{i,j}$ units. This design strategy minimizes reconfiguration overhead and optical weight storage, enabling compact and energy-efficient implementation on chip.

Figure 1(d) further illustrates several common matrix structures (e.g., block-circulant, Hadamard) that can be encoded through different configurations of the fixed unitary matrices. These serve as building blocks for deeper models and more complex transformations.

Biomedical Optics EXPRESS

Importantly, this decomposition enables MediONN to emulate layers of convolutional architectures such as U-Net, where each optical block corresponds to a small matrix acting on a local region or feature channel. While digital hardware can implement large convolutional kernels directly, our photonic chip is constrained to fixed-size computing units (e.g., 4×4). To bridge this gap, we partition the larger matrix operation into multiple smaller photonic-compatible submatrices, each realized optically. This allows scalable implementation of convolution-like behavior using repeated application of photonic 4×4 MVM blocks, aligned with techniques such as im2col unfolding.

Together, the four figures in Fig. 1 describe a tightly integrated hardware-software loop: the photonic hardware is first characterized and emulated via the DPE (figure a); the model is decomposed into optical-compatible MVMs (figure b); the matrix blocks are mapped onto the photonic chip using $B\Sigma P$ (figure c); and these units collectively represent structured, scalable transforms (figure d) used to construct layers of a full neural network.

Figure 2 shows the silicon photonic–electronic neural chip. In Fig. 2(a), the packaged 4×4 photonic neural computing chip is presented. The corresponding structured photonic circuit for matrix processing when $k = 4$ is illustrated in Fig. 2(b). In this layout, each optical input (at different wavelengths) is processed sequentially through three functional stages: an input unitary matrix P, a diagonal modulation matrix Σ, and an output unitary matrix B, together forming the structure $W_{i,j} = B\Sigma_{i,j}P$.

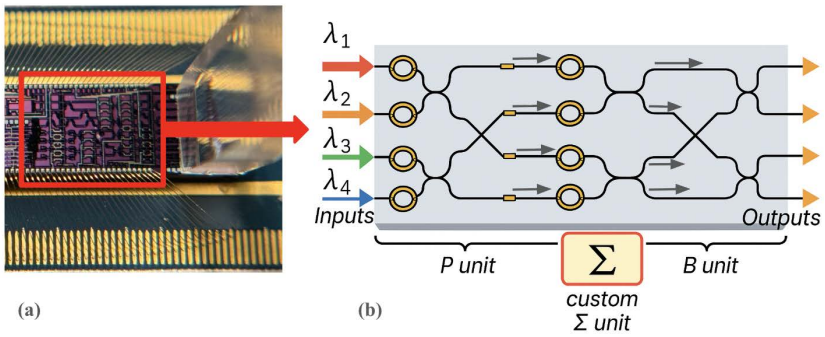


Fig. 2. (a) Packaged 4×4 photonic neural computing chip, and (b) structured photonic circuit for matrix processing when $k = 4$.

The unitary matrices P and B are implemented using passive butterfly-style mesh networks composed of directional couplers and phase shifters. These meshes realize fixed transforms—such as FFT, IFFT, or Hadamard matrices—depending on the configured internal phases, and are not trained during learning. The central diagonal block $\Sigma_{i,j}$ is constructed using two vertical columns of MZI-based optical attenuators, which adjust both the amplitude and phase of each optical signal. These MZIs, located in the diagonal (Σ) block, are the only trainable elements within the optical matrix-vector multiplication path. They are configured according to learned model weights mapped from the DPE-trained neural network.

To provide more details of the physical implementation, Fig. 3 further shows the complete design layout of the chip, with color-coded regions highlighting the functional blocks, including the input waveguides, fixed unitary mesh regions (P and B), and the trainable diagonal region (Σ). This structured decomposition enables low-latency and modular optical matrix-vector multiplication with minimal optical components, supporting scalable deployment of MediONN using cascaded 4×4 subblocks.

To enhance the performance and training stability of optical neural networks in medical image segmentation, we introduced a Gaussian-based weight initialization scheme tailored for photonic architectures. Our method employs zero-mean Gaussian distributions to initialize the optical

Biomedical Optics EXPRESS

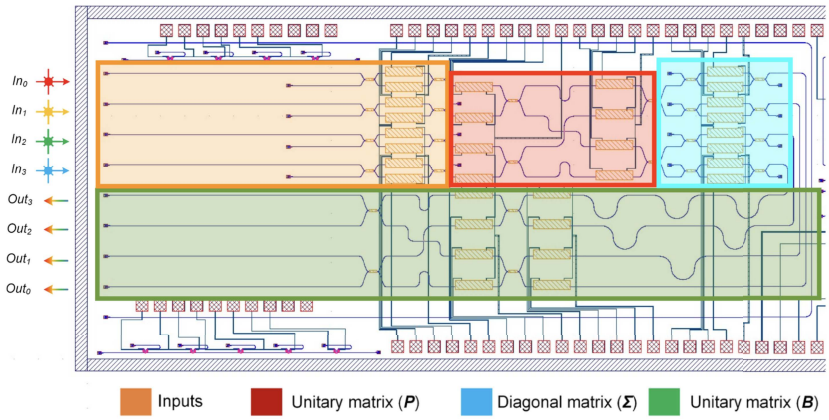


Fig. 3. Detailed design layout of the photonic chip shown in Fig. 2, with color-coded regions corresponding to the functional blocks.

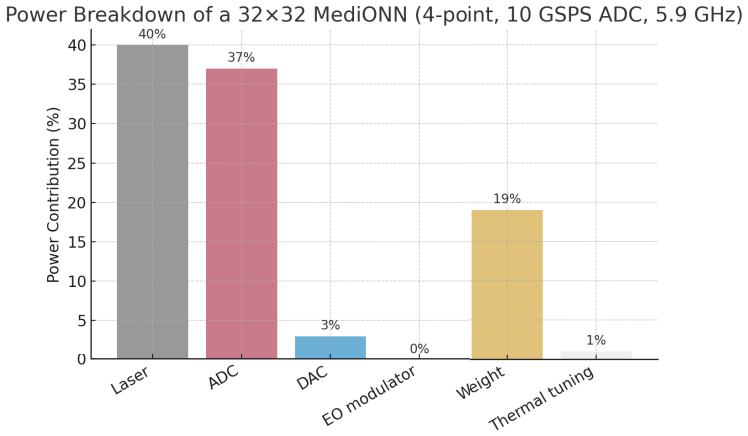


Fig. 4. Power breakdown of a 32×32 MediONN design operating at 5.9 GHz, showing contributions from modulators, phase shifters, ADCs, laser sources, and control electronics. The total system power is 3.3 W.

weight matrices, which better align with the statistical properties of medical imaging data and facilitate smoother gradient flow during training. In CT-based imaging, a large portion of pixels correspond to dark background regions with intensities near zero. Under Gaussian initialization, values around zero lie near the peak of the distribution, resulting in relatively large effective weights and stronger gradient signals. In contrast, functions such as *tanh* map values near zero back to small outputs, further shrinking already low-intensity signals and accelerating gradient vanishing. By aligning the initialization with the characteristics of CT data, this strategy provides a stable foundation for network training and contributes to more dependable feature extraction in the early stages.

In addition to the proposed Gaussian-based scheme, we also considered other possible initialization strategies [14,18]. For example, *tanh*-based initialization was experimentally tested, but in our early runs the training process diverged immediately, producing NaN (Not-a-Number) outputs. Further inspection revealed that many weights quickly collapsed to zero, leading to vanishing gradients and preventing convergence. We also examined the feasibility of fixed all-ones initialization and purely random initialization from a theoretical standpoint. In photonic

Biomedical Optics EXPRESS

architectures, initializing all weights to one forces identical optical paths at the outset, which removes the diversity needed for effective interference patterns. This lack of variation causes the network to behave as if all channels were coupled identically, resulting in vanishing gradients and severely limiting the network's capacity to learn distinct features. Conversely, purely random initialization produces highly irregular and uncorrelated phase distributions across the interferometers. Because optical signals combine through interference, such unstructured randomness often leads to strong destructive or constructive interference in arbitrary locations. When multiple layers are cascaded, these effects accumulate, amplifying oscillations in the training dynamics and frequently driving the system into unstable regimes. These considerations suggest that such approaches are not well suited for photonic implementations. By contrast, the proposed zero-mean Gaussian initialization balances weight diversity without introducing excessive phase irregularity, thereby avoiding both collapse and instability. This results in smoother gradient flow and consistently stable convergence in segmentation tasks.

We further expanded the system's capability by transitioning from 2D to 3D optical convolution, enabling volumetric feature learning across CT scan slices. While 2D photonic architectures have demonstrated efficiency in simple classification tasks, volumetric segmentation demands spatial coherence along the depth axis and significantly more complex hierarchical feature aggregation. To address this, we designed 3D optical convolutional layers within our photonic framework using silicon photonic chips, ensuring that light-based signal propagation supports voxel-level resolution. This architectural extension represents one of the first implementations of volumetric photonic computation for medical segmentation and paves the way for scalable, energy-efficient solutions in clinical imaging.

3. Dataset and data preprocessing

The NIH Pancreas-CT dataset is collected by the National Institutes of Health Clinical Center [19], this public dataset includes 81 contrast-enhanced abdominal CT scans, each with a resolution of 512×512 pixels. The number of slices varies from 181 to 466, with slice thicknesses between 0.5 and 1.0 mm. The dataset includes a diverse group of patients, consisting of 53 men and 27 women, with ages ranging from 18 to 76 years and an average age of 46.8 years. A medical student performed the detailed segmentation of the pancreas slice-by-slice, and an experienced radiologist verified and refined this work, ensuring the accuracy of the labels that are essential for the research. The annotated masks, provided in Neuroimaging Informatics Technology Initiative (NIfTI) format, denote areas with anomalies using pixel values of 1, while regions without anomalies are marked with a value of 0.

The dataset is utilized for both 2D and 3D medical image segmentation, where each slice is considered independently in the 2D approach, while the 3D approach analyzes the volume constructed from a collection of slices for each patient. Patients are randomly assigned to training, validation, and test sets using a 70/10/20 split, yielding 56, 7, and 17 pairs of images and masks, respectively. The training data is shuffled prior to batching to ensure varied input. Given the repetitive structures in three-dimensional medical images, each patient's image and mask are resized to dimensions of $128 \times 256 \times 256$ pixels, where the first dimension corresponds to the total number of slices and the subsequent dimensions represent the size of each 2D slice.

Due to GPU memory constraints, it is not feasible to use the entire volume of each patient in a single batch for the 3D study; therefore, each patient's image and mask are divided into smaller volumes, referred to as patches. The size of these patches serves as a hyperparameter in this paper. To prevent overfitting, data augmentation is applied to the training data. For the 2D segmentation, augmentation techniques such as rotation, vertical flipping, padding, and grid distortion are employed. In contrast, the 3D study utilizes transforms including 3D random affine transformation, random elastic deformation, and random flip, specifically vertical flips, to enhance the generalization of the model.

Biomedical Optics EXPRESS

4. Experiments

The segmentation tasks are performed using 2D-UNet [20] and 3D-UNet [21], both fully convolutional networks that excel in biomedical image segmentation. We construct these models within our MediONN framework. To implement the models, we utilize MVM operations in conjunction with the widely applied im2col tensor unrolling method to realize convolution [22]. Large tensor operations are partitioned into 4×4 blocks and efficiently mapped onto our MediONN architecture. In order to reuse the convolution and reduce computational complexity in the upsampling step, we approximate transpose convolution using interpolation techniques, such as bilinear interpolation. These interpolation methods provide a more efficient way of upsampling feature maps, allowing for faster processing while still achieving acceptable upsampling results. By offering a smooth upsampling operation, they help preserve the quality of the feature maps during upscaling. To ensure consistency between both the optical (MediONN) and electronic models, the same approximation of transpose convolution using interpolation techniques is applied to the electronic model.

Before finalizing the training pipeline, we encountered a critical issue during early-stage experiments. Initially, we used the common `tanh` activation function in optical training. However, the network consistently failed to generate meaningful predictions. Upon closer investigation, we discovered that many weights had collapsed to zero, revealing a severe gradient vanishing problem.

This issue can be understood by comparing the properties of activation functions. The `tanh` function outputs zero when its input is zero, whereas a Gaussian-like function reaches its peak under the same condition. In the context of medical imaging, input data are often sparse and dark, with many pixel values near or equal to zero. Consequently, when such inputs are passed through `tanh`, the resulting activations approach zero, effectively suppressing gradient flow.

To overcome this, we employed Gaussian initialization to encourage better gradient propagation in the early training phase. This adjustment significantly improved stability and prevented the collapse of weights. While initialization is a well-established topic in deep learning, its interaction with data sparsity in optical systems has not been widely addressed. Our results suggest that weight initialization should be tailored to the distribution of input data—particularly in hybrid architectures—highlighting a broader principle in training strategy design.

Central to our approach is the use of a neural network–based estimator that accurately models the photonic integrated circuit (PIC)'s non-ideal behaviors during both forward propagation and gradient backpropagation. By explicitly capturing these physical chip characteristics, this estimator enables gradient-based optimization that is cognizant of inherent variations, thereby facilitating more effective and variation-aware training.

The 2D-UNet architecture consists of an encoder and a decoder path, each containing four resolution steps. The encoder path captures important information from the image, while the decoder path upsamples the feature maps from the encoding path and constructs the final segmentation mask. Each path contains four blocks, with each block comprising two 3×3 convolutions followed by batch normalization to accelerate convergence and a Sigmoid-weighted Linear Unit (SiLU) activation. Additionally, a 2×2 max pooling layer with strides of two is applied in each dimension.

In the decoder path, each layer begins with an upconvolution of 2×2 with strides of two, which is then followed by two 3×3 convolutions along with a ReLU activation. Notably, shortcut connections from layers of equal resolution in the encoder play a crucial role by providing high-resolution features to the decoder, enhancing the overall segmentation accuracy. To effectively prevent bottlenecks in the network, the number of feature channels is doubled after each max pooling operation and halved during each upconvolution. Additionally, dropout is implemented with a rate of 20% after each max pooling and upconvolution layer to further prevent the risk of overfitting.

Biomedical Optics EXPRESS

Ultimately, the final layer employs a 1×1 convolution to reduce the number of output channels to one, and a sigmoid function is applied at each pixel to compute the loss. Weight initialization is strategically conducted using a Gaussian distribution with a standard deviation of $\sqrt{\frac{2}{N}}$, where N represents the number of incoming nodes for each neuron. Furthermore, the 3D-UNet architecture extends the capabilities of the 2D-UNet by replacing all 2D operations with their 3D equivalents, encompassing 3D convolutions, 3D max pooling, and 3D upconvolutional layers. All experiments are conducted in a PyTorch environment on NVIDIA A100 GPUs.

In medical image segmentation, datasets often contain an imbalance, with a greater number of anomaly voxels compared to those without anomalies. This imbalance can lead to segmentation predictions that prioritize high precision—correctly identifying positive cases—while often missing actual anomalies, resulting in low sensitivity (recall). Consequently, the model may produce fewer false positives but also overlook many true positives, a situation that is not ideal for computer-aided diagnosis and clinical decision support systems. To address this challenge and achieve a more favorable balance between precision and sensitivity, Salehi et al. [23] proposed a loss function based on the Tversky index, which generalizes both the DSC and the F_β score.

In our implementation, we adopt the Tversky loss, which is closely related to the DSC and F_β score. The detailed definitions of these metrics are provided in Section 6, where we evaluate the model performance.

The Tversky loss function, formulated as:

$$\text{Loss} = \frac{TP}{TP + \alpha \cdot FP + \beta \cdot FN} \tag{1}$$

This loss function is employed in this work to enable a flexible trade-off between false positives and false negatives.

For a comprehensive quantitative evaluation of the network's performance, metrics such as specificity, sensitivity, precision, F1 score, F2 score, and DSC are calculated and reported.

The comparison between the network output and ground truth labels is conducted using sigmoid nonlinearities in conjunction with the Tversky loss function. Specifically, the sigmoid function transforms the network's raw output into probability scores between 0 and 1, allowing voxels with probabilities of 0.5 or greater to be classified as having anomalies, while those below this threshold are considered normal. Furthermore, all models are trained with parameters set to $\alpha = 0.3$ and $\beta = 0.7$, where α and β control the sensitivity of the Tversky loss function to false positives and false negatives, respectively. To find the optimized model, the training process utilizes the Adam optimizer to minimize the Tversky loss. This iterative optimization aims to adjust the model parameters effectively, enhancing performance by reducing both false positives and false negatives. The training schedule follows Leslie Smith's one-cycle learning rate policy [24], spanning 100 epochs. In this approach, the learning rate first increases from a lower value to a peak rate before decreasing back to the initial lower value, with both phases occurring in two equal-sized steps. Ultimately, the maximum learning rate is fine-tuned for each model, while the lower rate is approximately one-tenth of the maximum. This policy not only accelerates convergence by enabling efficient exploration of the loss landscape but also enhances generalization, thereby reducing the risk of overfitting. The optimal model is subsequently selected based on improvements in the DSC during validation.

5. Photonic chip characterization

To evaluate the feasibility of our MediONN system, we first characterize key device- and system-level parameters of the fabricated photonic chip, including optical insertion loss, phase stability, thermal tuning power consumption, and tolerance to static errors. These metrics not only validate the practicality of the hardware but also provide a basis for comparison with electronic

Biomedical Optics EXPRESS

accelerators. By situating our results in the broader context of integrated photonics, we ensure that the reported performance aligns with established device benchmarks while highlighting MediONN's unique system-level advantages. [12,14,25–29]

5.1. Optical insertion loss

To assess the feasibility of end-to-end photonic computation, we characterized the optical insertion loss of our fabricated chip at the component level. The total loss in the MZI-based matrix–vector multiplication (MVM) network primarily arises from directional couplers, phase shifters, and waveguide crossings—each contributing to attenuation along the optical path.

In our current 4×4 photonic matrix block design, the longest optical path traverses six 2×2 directional couplers, four phase shifters, and two waveguide crossings. As summarized in Table 1, based on typical insertion loss values at 1550 nm—approximately 0.3 dB per coupler, 0.2 dB per phase shifter, and 0.1 dB per crossing—the cumulative insertion loss is estimated to be around 3.2 dB per matrix block.

Table 1. Typical insertion loss per optical component (values adapted from AMF PDK [30]).

Optical Component	Insertion Loss (dB)
50/50 directional coupler	0.3
Phase shifter	0.2
Waveguide crossing	0.1

To analyze scalability, we consider the cascading of multiple matrix blocks in a deep neural network. Assuming an input optical power of 0 dBm and a photodetector sensitivity threshold of –20 dBm, the system can tolerate up to approximately 6 sequential matrix blocks before the signal falls below the detection threshold. (This estimate assumes 3.2 dB loss per block: $20 \div 3.2 \approx 6.25$.) This estimate does not account for additional losses from routing or packaging, so practical implementations may require fewer stages or the inclusion of optical amplifiers.

Importantly, our architecture supports wavelength-division multiplexing (WDM), which enables parallel matrix–vector multiplications across multiple wavelengths. This parallelism effectively distributes the optical power budget and mitigates the impact of insertion loss by reducing the number of sequential stages required per wavelength. For example, using 4 wavelengths to process different segments of the computation allows the system to maintain throughput while limiting the depth of each optical path.

Additionally, all optical paths were carefully balanced during layout to minimize differential attenuation, and an on-chip calibration procedure was implemented to compensate for residual imbalances caused by fabrication variations or thermal drift. These measures help preserve inference accuracy even as the system scales.

This loss-aware design ensures robust performance within the system's operating constraints. While future iterations may benefit from emerging low-loss photonic components, our current prototype demonstrates acceptable scalability for practical deep learning tasks.

As a feasibility analysis, insertion-loss values were taken from the Advanced Micro Foundry (AMF) Process Design Kit (PDK) [30]; the cumulative path loss was then estimated by counting optical elements along the longest signal path and cross-checked against our component-level characterizations for consistency.

5.2. Phase stability and end-to-end online reasoning

In photonic neural networks, phase stability is a critical concern due to the sensitivity of optical interference to environmental variations and fabrication imperfections. To ensure robust

Biomedical Optics EXPRESS

computation, our system incorporates an on-chip calibration mechanism that characterizes and compensates for phase drift and static bias in the photonic devices.

Each phase shifter in our MediONN exhibits a unique tuning response, modeled as a phase shift $\phi(V) = \alpha V^2 + \phi_0$, where α is the tuning factor and ϕ_0 is a static phase offset resulting from fabrication variations. These parameters are experimentally calibrated by sweeping the input voltages and fitting the resulting transmission curve.

For phase shifters embedded within MZIs, the calibration is extended to estimate the differential phase bias between the two arms. Each MZI attenuator was tuned by scanning the applied heater voltages across its operating range while recording the corresponding transmission spectrum. The measured data were then fitted to a quadratic response, enabling extraction of both the tuning factor α and the static offset ϕ_0. This direct measurement ensures that device-specific nonidealities are accurately incorporated into subsequent system-level training and inference. The calibrated values for the Σ-unit attenuators are summarized in Table 2.

Table 2. Calibrated parameters of MZI attenuators in the Σ unit.

MZI Attenuator	Tuning Factor α (rad/V^2)	Phase Offset ϕ_0 (rad)
$\Sigma_{1,1}$	1.896	0.32
$\Sigma_{1,2}$	1.855	0.24
$\Sigma_{2,1}$	1.929	−0.67
$\Sigma_{2,2}$	1.874	−1.34

To further enhance phase robustness and enable end-to-end reasoning, we leverage the DPE, trained on chip-level I/O measurements, as a neural-network–based surrogate model that captures static phase offsets and nonlinear tuning characteristics of the fabricated chip.

During training, the DPE operates in a *differentiable mode*, acting as a smooth, gradient-compatible function that links control voltages to predicted optical outputs. This allows backpropagation to account for device-specific imperfections—such as phase offsets, nonlinear tuning curves, and thermal drift—so that the learned photonic weights are already adapted to the physical hardware constraints.

For validation and deployment, the DPE switches to a *lookup mode*, where it directly queries a database of experimentally measured voltage–intensity pairs to reproduce the chip's actual responses. This ensures that inference reflects the true behavior of the device rather than an idealized model.

Feasibility and Bottlenecks of Online Reasoning: This dual-mode operation supports end-to-end online reasoning by enabling real-time inference (i.e., inference without retraining or re-optimization) through direct voltage-to-output mapping. However, several bottlenecks must be considered:

- *Lookup Latency:* While the lookup mode avoids computational overhead from gradient evaluation, querying large voltage–intensity tables can introduce latency, especially if the resolution of control voltages is high.
- *Thermal Drift and Recalibration:* Environmental fluctuations may cause phase drift over time, requiring periodic recalibration or adaptive compensation mechanisms to maintain inference fidelity.
- *Memory Overhead:* Storing high-resolution lookup tables for all photonic components can be memory-intensive, particularly for large-scale systems with many phase shifters.
- *Scalability Limits:* As the system scales to deeper networks or more wavelengths, the complexity of maintaining synchronized calibration across all channels increases, potentially impacting throughput. Modular architectures and wavelength-division multiplexing

Biomedical Optics EXPRESS

(WDM) schemes, discussed elsewhere in this paper, can help mitigate these scalability challenges.

Despite these challenges, our architecture demonstrates that end-to-end online reasoning is feasible with careful calibration and DPE-guided training. The lookup-based inference ensures hardware-faithful execution, while the differentiable mode enables robust optimization under real-world constraints.

5.3. System-level power analysis and thermal tuning characteristics

Thermal phase tuning is a key contributor to the overall power budget of photonic neural networks. In our MediONN prototype, each thermo-optic phase shifter requires less than 2.5 mW to achieve a full π phase shift at 1550 nm, consistent with reported values for silicon photonic devices. The chip was fabricated by Advanced Micro Foundry (AMF) and tested using a Printed Circuit Board (PCB)-integrated setup with a 40-channel Digital-to-Analog Converter (DAC) controlled by a Raspberry Pi 4. For high-speed digitization, a 10 GSPS (giga-samples per second) Analog-to-Digital Converter (ADC) was employed, which defines the electronic I/O power contribution in the system.

At the system level, we evaluate a representative 32×32 MediONN design operating at 5.9 GHz (Fig. 4). This larger configuration is constructed by tiling multiple 4×4 photonic matrix blocks, reflecting the scalable architecture used in deep neural network layers. The total power consumption of such a system is estimated at 3.3 W, including contributions from electro-optic modulators, thermal phase shifters, ADCs, laser sources, and control electronics. Power values were obtained through a hybrid methodology: the consumption of devices such as DACs, ADCs, and control circuits was taken directly from experimental measurements or vendor datasheets, while the overall system-level total was extrapolated according to the number of components required in a 32×32 layout. This combination of direct measurement and specification-based extrapolation ensures that the reported numbers are both hardware-faithful and scalable, providing a reliable basis for evaluating thermal tuning power consumption in MediONN.

5.4. Comparison with electronic accelerators

To contextualize the efficiency of MediONN, we compare its hardware-level metrics with those of conventional electronic accelerators such as GPUs and TPUs. MediONN's performance metrics are derived from experimentally validated device parameters, including electro-optic modulator speed, photodetector response, ADC/DAC delay, and thermal tuning power. These metrics are independent of the training dataset and instead reflect the physical characteristics of the hardware.

In terms of latency, MediONN achieves an effective MVM delay of ~164 ps, corresponding to an operating frequency of ~6 GHz. This delay is orders of magnitude lower than the millisecond-scale latency typical of GPUs and TPUs when executing the same matrix operations.

Energy efficiency is another dimension where MediONN demonstrates clear advantages. Based on measured and modeled component power, MediONN reaches ~9.5 Tera-Operations Per Second per Watt (TOPS/W), compared to ~2.3 TOPS/W for the Google TPU and only ~0.19 TOPS/W for the Nvidia Tesla P40 GPU. While memristor arrays can achieve higher energy efficiency (~28 TOPS/W), they typically operate at much lower clock speeds than photonic implementations.

Finally, MediONN achieves a computational density of up to ~225 TOPS/mm^2, far exceeding that of electronic accelerators (e.g., ~0.28 TOPS/mm^2 for TPUs). This high density arises from the inherent parallelism of Wavelength-Division Multiplexing (WDM) and the compact footprint of silicon photonics.

As summarized in Table 3, these comparisons demonstrate that MediONN outperforms conventional electronic accelerators in both latency and energy efficiency, while also achieving a significantly higher compute density.

Biomedical Optics EXPRESS

Table 3. Comparison of MediONN with representative electronic accelerators. Metrics are derived from experimentally validated device parameters.

Platform	Inference Delay	Energy Efficiency (TOPS/W)	Compute Density (TOPS/mm^2)
MediONN (Photonic)	~164 ps (~6 GHz)	9.5	225
Google TPU	ms-level	2.3	0.28
Nvidia Tesla P40 GPU	ms-level	0.19	0.1
Memristor Array	μs-level	28	N/A

Taken together, these results highlight the inherent optical advantages of MediONN in both latency and energy efficiency, reinforcing its potential for scalable and high-performance biomedical computing applications.

5.5. Static error compensation

Fabrication variations in photonic components—such as unbalanced splitting ratios in directional couplers and static phase offsets in interferometers—can significantly degrade the accuracy of optical neural networks. To address these static errors, MediONN incorporates multiple layers of compensation.

First, on-chip calibration is performed by sweeping the control voltages of phase shifters and MZI attenuators to extract device-specific tuning factors and phase offsets. These calibrated parameters are then used to correct the control voltages applied during inference.

Second, we integrate a DPE, trained on experimentally measured I/O responses, into the MediONN training loop. The DPE captures static offsets and nonlinear tuning curves, ensuring that the optimization is aware of device-level nonidealities.

Finally, hardware-aware training, combined with initialization strategies that avoid extreme operating points, ensures that the photonic weights remain robust under fabrication-induced perturbations. Together, these strategies mitigate the impact of static fabrication errors and preserve stable inference accuracy. Our methodology is also conceptually aligned with recent efforts in error-tolerant photonic optimization and scalable configuration methods [31,32], which emphasize the necessity of adaptive calibration and robust optimization strategies for practical ONN deployment.

5.6. Wavelength-division multiplexing (WDM) compatibility

To enhance system-level throughput and parallelism, MediONN is designed to support wavelength-division multiplexing (WDM). In our prototype implementation, input signals corresponding to different data channels are modulated onto distinct optical carriers (e.g., $\lambda_1 = 1548.0$ nm, $\lambda_2 = 1549.0$ nm, etc.) using resonator-based electro-optic modulators. These multi-wavelength signals are then coupled into the chip via broadband grating couplers that support simultaneous injection of multiple wavelengths into the same optical path.

Inside the chip, each wavelength traverses identical photonic submatrix blocks, enabling parallel matrix–vector multiplications without requiring phase coherence between channels. Since each wavelength experiences the same optical path and layout, the system exhibits uniform response across channels, with minimal inter-wavelength crosstalk due to the inherent spectral selectivity of the modulators and routing elements.

At the output stage, wavelength-resolved photodetection allows each channel to be independently measured without requiring phase-sensitive detection. This simplifies the readout circuitry and enables robust passive multi-channel operation, supporting scalability across wavelength channels.

Biomedical Optics EXPRESS

6. Results and discussion

Evaluation Metrics Each model is evaluated using the following metrics:

- **Specificity:** The proportion of true negatives correctly identified. It measures how well the model identifies **negative cases**.
- **Sensitivity (Recall):** The proportion of true positives correctly identified. It measures how well the model identifies **positive cases**.
- **Precision:** The proportion of true positives among all positive predictions. It measures the **accuracy** of the positive predictions.
- **F1 Score:** The harmonic mean of precision and sensitivity. It ensures a good **balance** between the two metrics, especially in cases of class imbalance.
- **F2 Score:** Similar to the F1 score but places more emphasis on sensitivity (recall) than precision. This is useful in cases where **false negatives** are more critical.
- **Dice Similarity Coefficient (DSC):** A commonly used metric in medical image segmentation. It is defined as:

$$DSC = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

- **F_β Score:** A generalization of the F1 score that weights recall β times more than precision. Defined as:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\beta^2 \cdot \mathrm{Precision} + \mathrm{Recall}} \tag{3}$$

  In this formula, when $\beta = 1$, it calculates the F_1 score, which gives equal weight to precision and recall.

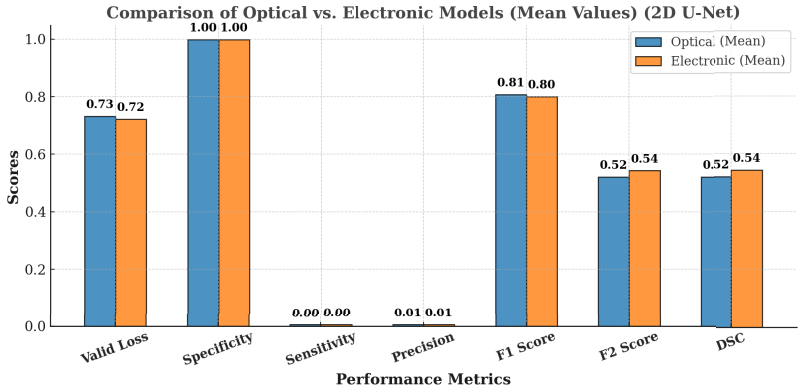
6.1. 2D model results with MediONN

Figure 5 presents the experimental mean and maximum results for the 2D pancreatic cancer segmentation task, using MediONN and conventional electronic models. The original data are provided in Supplement 1. The scores reported include mean and maximum values for each metric, such as DSC, F1 Score, and others. Notably, the DSC score for the optical (MediONN) model means 0.5215, while the electronic model means 0.5432. They reach a similar degree of performance. When considering the maximum results, both models achieve 0.5919 for DSC, indicating that the performance at peak values is equal for both models.

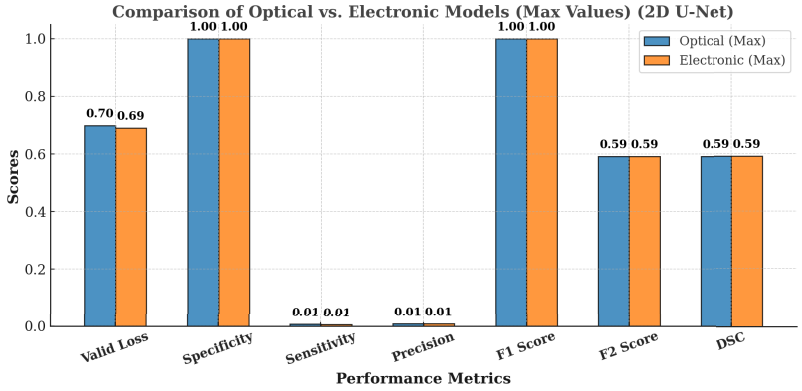
In terms of other key metrics, the optical model shows almost equal performance to the electronic model in mean F1 Score (0.8076 vs. 0.8007) and F2 Score (0.5213 vs. 0.5431), with slight variations that do not suggest a significant difference in performance. Both models exhibit high specificity, indicating their effectiveness in correctly identifying negative cases, but their sensitivity remains low, suggesting that they tend to miss some positive cases, which is consistent with the characteristics of the dataset.

The optical model demonstrates comparable performance in key metrics and offers practical advantages such as lower latency and reduced energy consumption, making it a promising alternative for real-world applications.

Biomedical Optics EXPRESS



(a) Mean results



(b) Maximum results

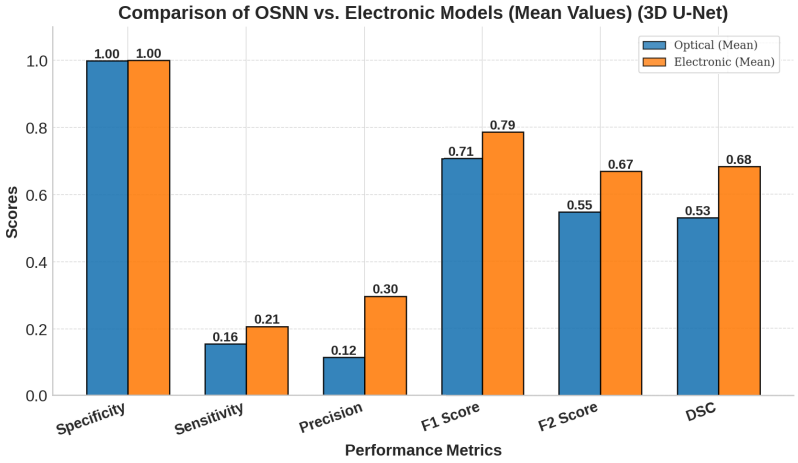
Fig. 5. Performance metrics of 2D U-Net models for optical (MediONN) and electronic. (a) Mean results, and (b) Maximum results. MediONN results are represented in blue, while electronic model results are shown in orange. Higher values for specificity, precision, F1 score, and F2 score are preferred; a score of 0 indicates the worst performance, while 1 signifies perfection. Conversely, for validation loss, lower values indicate better performance.

6.2. 3D model results with MediONN

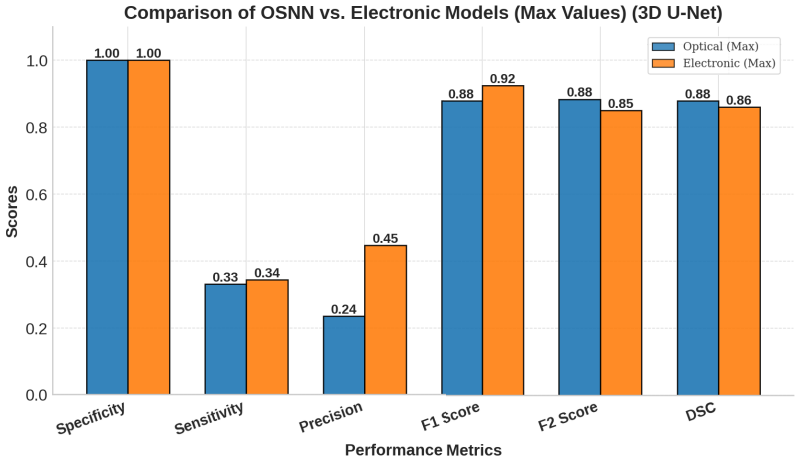
Figure 6 presents the mean and maximum performance metrics for the model trained with both MediONN and electronic chips for comparison. The mean DSC for the 3D MediONN model is 0.5302, while the 3D electronic model achieves 0.6821. However, in terms of maximum DSC, MediONN reaches 0.8788, which is approximately equivalent to the electronic model's 0.8599, demonstrating that MediONN has the potential to achieve equally strong segmentation performance in optimal conditions.

In addition to DSC, the comparison extends to other key metrics. The mean F1 Score for MediONN is 0.7066, slightly lower than the 0.7850 recorded for the electronic model, suggesting that MediONN still maintains a competitive balance between precision and recall. A similar pattern is observed in the F2 Score, where MediONN achieves 0.5470, compared to 0.6678 for the electronic model. While the electronic model shows a stronger recall tendency, MediONN's results indicate a more stable performance across different conditions.

Biomedical Optics EXPRESS



(a) Mean results



(b) Maximum results

Fig. 6. Performance metrics of 3D U-Net models for optical (MediONN) and electronic. (a) Mean results, and (b) Maximum results. MediONN results are represented in blue, while electronic model results are shown in orange. Higher values for specificity, precision, F1 score, and F2 score are preferred; a score of 0 indicates the worst performance, while 1 signifies perfection. Conversely, for validation loss, lower values indicate better performance.

Importantly, both models exhibit exceptionally high specificity, with MediONN averaging 0.9979, nearly matching the electronic model's 0.9995. This underscores MediONN's strong ability to correctly identify negative cases, which is crucial for applications requiring minimal false positives. While the electronic model achieves a mean sensitivity of 0.2074, compared to 0.1559 for MediONN, this difference is a trade-off often seen in high-specificity models.

While MediONN achieves performance comparable to electronic models in 2D tasks, its 3D results show a relative decline, motivating an analysis of possible contributing factors. We hypothesize that the performance drop observed in the 3D segmentation task is primarily due to cumulative approximation errors introduced by the optical convolutional layers. In our current prototype, each optical matrix multiplication is simulated using a DPE, which is trained on

Biomedical Optics EXPRESS

real-world photonic hardware measurements to approximate the behavior of the optical chip. Unlike numerical computation in electronic models (which use exact floating-point operations), the DPE introduces small prediction errors in each layer.

These errors, while relatively insignificant in shallow 2D architectures, can accumulate and potentially amplify—possibly in a nonlinear fashion—as the network depth and complexity increases in 3D segmentation tasks. In addition, optical convolution operations may involve additional patch-based approximations or block-wise inference (e.g., fixed 4×4 kernels), which further compound modeling discrepancies. This can explain the observed DSC gap compared to the electronic baseline.

We plan to mitigate this issue in future work by refining the DPE model, improving optical chip calibration, and exploring hybrid architectures where optical components handle high-throughput layers while maintaining critical precision paths electronically.

Furthermore, comparing the results in Fig. 6 to those in Fig. 5, the transition from 2D to 3D U-Net models provides substantial advantages in capturing volumetric data and spatial relationships, which are essential in medical imaging applications. Notably, all the evaluation metrics show improvement in the 3D model, highlighting its enhanced performance over the 2D model.

Figure 7 illustrates the data distribution using a violin plot. A violin plot visualizes both the distribution and summary statistics of the data. The width of the shape indicates the density of values at different points, showing how frequently they occur. The central dash line represents the median (50th percentile), while the upper and lower dash lines correspond to the 75th and 25th percentiles, respectively. This makes violin plots particularly useful for understanding both the spread and overall distribution of the data.

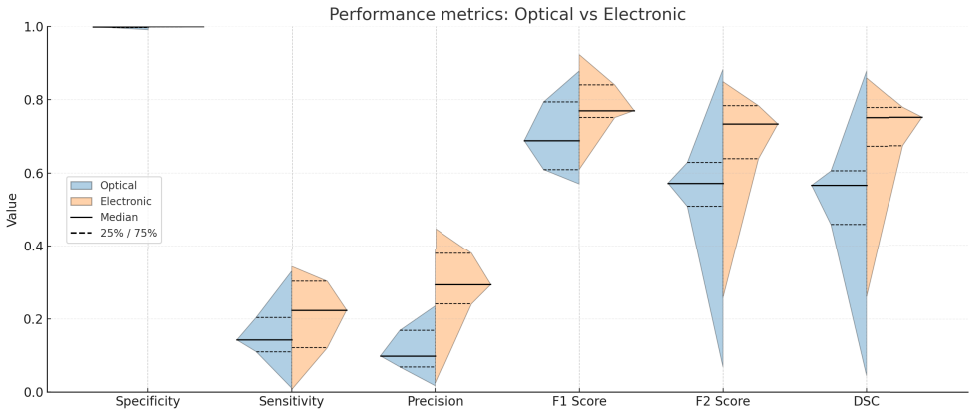


Fig. 7. Violin plot for the performance metrics for 3D U-Net models with MediONN and electronic. MediONN results are represented in blue, while electronic model results are shown in orange. Higher specificity, precision, F1 score, and F2 score indicate superior performance, with scores approaching 1 signifying optimal outcomes. Conversely, lower validation loss values correspond to better model efficacy.

Using this violin plot structure, we interpret our test results with 3D U-Net, as shown in Fig. 7, where we apply our model, trained with MediONN, to a test dataset of 17 patients. This figure also includes results from the electronic model for comparison. By utilizing the violin plot, we gain a deeper understanding of how performance metrics are distributed across the patient dataset, revealing key patterns and potential variations beyond simple summary statistics.

To better assess the distribution of these metrics, we examine key percentiles:

- The **25th percentile** indicates that 25% of patients' scores fall below this value.

Biomedical Optics EXPRESS

- The **50th percentile (median)** represents the middle of the data, with half of the patients scoring below and half above.
- The **75th percentile** indicates that 75% of patients' scores fall below this value, and the remaining 25% have scores above it.

The original data can be found in Supplement 1.

These percentiles are crucial not only for summarizing the central tendency (mean and median) but also for providing insights into the data's distribution and variability. Specifically, the interquartile range (IQR), which represents the difference between the 25th and 75th percentiles, helps assess the spread of the middle 50% of the data. A narrower IQR suggests more consistent performance, while a wider IQR indicates greater variability. For the DSC, the IQR—calculated as the difference between the 75th percentile (0.6059) and the 25th percentile (0.4588)—is 0.1471 (data are in Supplement 1), suggesting moderate variability in the model's performance. While the model demonstrates a reasonable level of consistency, this variability indicates that its effectiveness may vary across different patients. In this case, the IQR reflects moderate variability, offering a more nuanced view of how the model generalizes across patients, beyond just the mean or maximum values.

Besides Fig. 7, which presents the overall statistical results, Fig. 8 provides a detailed breakdown of performance metrics for each patient in the test dataset (the original data are available in Supplement 1). This table evaluates the effectiveness of our trained 3D MediONN model in predicting clinical outcomes across 17 patients, offering valuable insights into its performance. The results indicate that the model consistently achieves high specificity, demonstrating strong capability in correctly identifying negative cases. The sensitivity values exhibit variability across patients, reflecting the challenges of detecting positive cases in certain instances. Additionally, the F1 and F2 scores remain relatively high, underscoring the model's ability to maintain a balance between recall and precision. These findings highlight the model's potential for delivering consistent and trustworthy predictions in clinical applications.

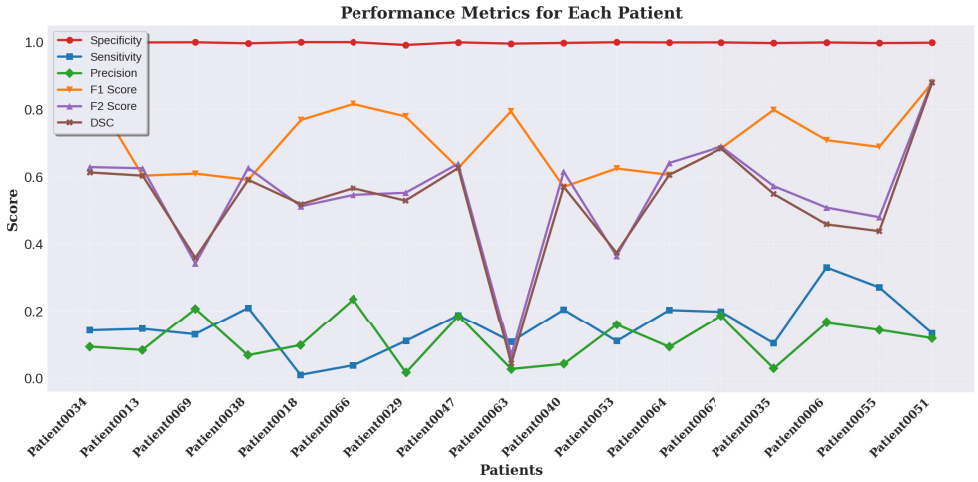


Fig. 8. Patient Performance Metrics with 3D MediONN: This table presents the evaluation metrics derived from applying our trained model to a test dataset of 17 patients. Each patient's performance metrics offer valuable insights into the model's effectiveness, revealing variations in specificity, sensitivity, precision, and F1 scores.

In Fig. 9, we visualize the patient predictions from our trained 3D MediONN by displaying CT scans, corresponding ground truth masks, and prediction outputs. The alignment between the

Biomedical Optics EXPRESS

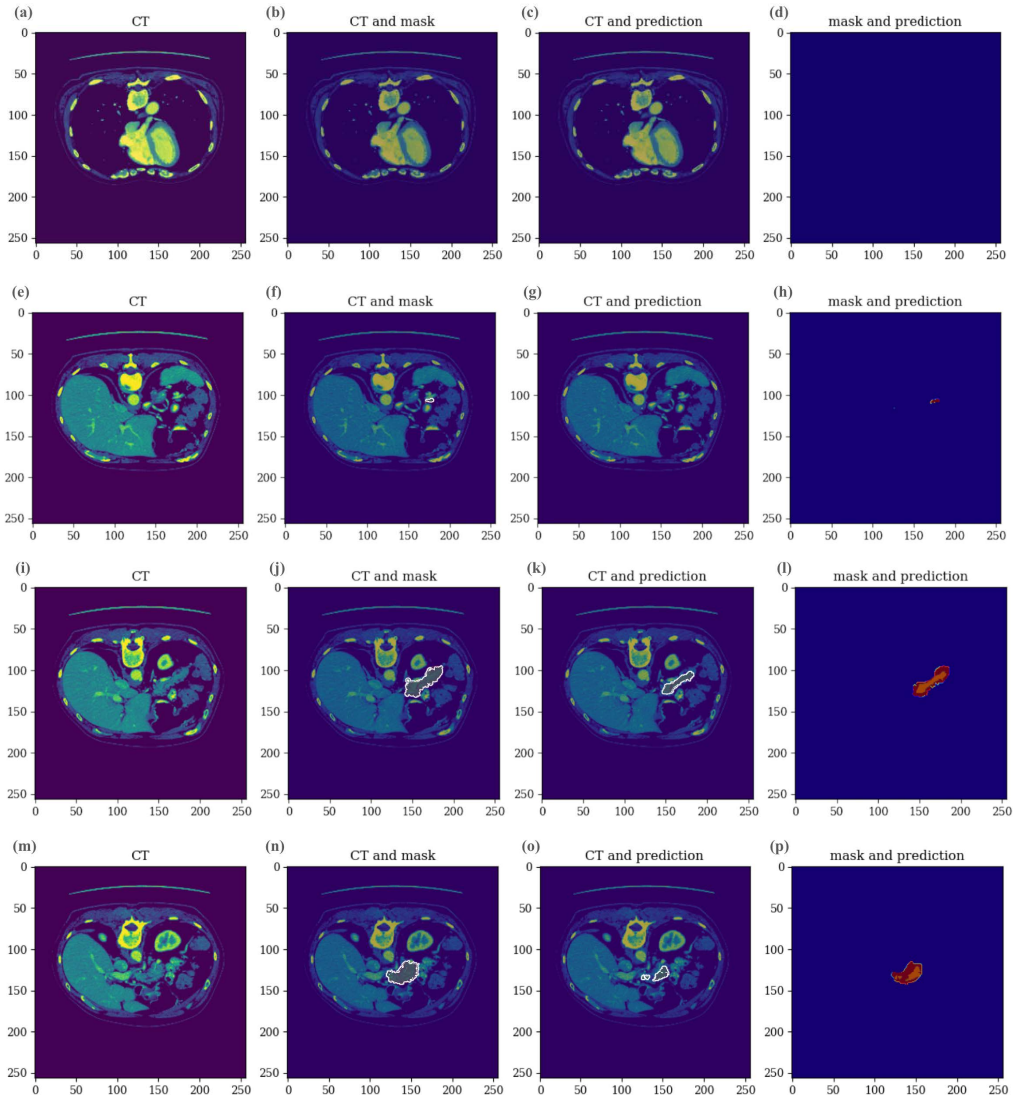


Fig. 9. Visualization of patient predictions with 3D MediONN. The figure follows the scan sequence from top to bottom, illustrating cases from the absence of tumors to their appearance. From left to right, each column represents the original CT scan (a, e, i, m), the ground truth mask overlaid on the CT scan (b, f, j, n), the predicted mask (c, g, k, o), and a comparison of both the ground truth and predicted masks (d, h, l, p).

ground truth masks and predictions is highly consistent; for instance, in cases where no tumor is present, both the mask and prediction accurately reflect this absence, highlighting the model's high specificity. The images follow the scanning order, demonstrating how the model responds as the CT scan transitions from an initial view of the area without a tumor to progressively closer scans, revealing the tumor as its features become increasingly discernible, while the prediction mask accurately identifies the tumor size and location throughout this process.

Taken together, these segmentation metrics constitute the end-to-end inference performance of our MediONN system on pancreatic CT segmentation, a substantially challenging pixel-level

Biomedical Optics EXPRESS

medical imaging task. The ability of MediONN to reproduce competitive Dice, precision, and sensitivity confirms that our optical pipeline operates as a full-system neural network, demonstrating its practicality and scalability for biomedical applications.

Furthermore, these results are not based on an idealized model but are inferred from the measured chip through our DPE. The DPE, trained on experimentally collected I/O data, captures static phase offsets and nonlinear tuning characteristics of the fabricated photonic circuit. By operating in differentiable mode during training and lookup mode during validation, the DPE ensures that the reported performance faithfully reflects the behavior of the real hardware.

7. Conclusion

We present a optical-electronic hybrid system, MediONN, specifically designed for the challenging task of 3D medical image segmentation. Unlike prior ONN research that primarily focused on image classification, MediONN addresses the unique computational demands of volumetric segmentation to increase the accuracy. The innovations include voxel-level prediction, spatial continuity, and multi-scale feature extraction—achieved by introducing a suite of photonic hardware innovations and algorithmic enhancements.

At the hardware level, we implement a packaged 4×4 MediONN photonic chip, embedded within a hierarchical 3D convolutional architecture. To overcome training instability commonly encountered in optical systems, we propose a Gaussian-based weight initialization scheme tailored for segmentation tasks, significantly improving convergence and model robustness. Additionally, we extend conventional 2D optical computing pipelines into 3D volumetric inference, enabling our photonic system to process full medical volumes with minimal latency and energy overhead.

On the NIH pancreas dataset, MediONN achieves competitive segmentation performance in both 2D and 3D settings. For 2D segmentation, it attains a DSC of 0.5215 and reaches a peak of 0.5919, alongside strong performance in other key metrics (e.g., specificity: 0.9986, F1 score: 0.8076). For 3D segmentation, MediONN attains a mean DSC of 0.5302, closely approaching the electronic baseline (0.6821), with maximum DSC reaching 0.8788 exceeding the baseline in several individual cases. These results confirm that photonic models can deliver reliable, high-fidelity segmentation outputs at a fraction of the energy cost.

By achieving high segmentation accuracy with reduced latency and power consumption, MediONN establishes a new milestone for optical neural networks in biomedical imaging. Our results demonstrate not only the feasibility but also the scalability of ONN-based architectures for complex pixel-wise tasks. This work paves the way for a new generation of photonic deep learning systems tailored to domain-specific, high-resolution medical applications.

Funding. Air Force Office of Scientific Research (FA9550-23-1-0452, FA9550-17-1-0071); National Institutes of Health (HHSN261201000085C).

Acknowledgments. The authors acknowledge support from the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR) (FA9550-17-1-0071) and from the project 'Towards Next-Generation Electronic-Photonic Devices' (FA9550-23-1-0452). This work was also supported in part by the National Institutes of Health under grant number HHSN261201000085C.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are available in the Supplement 1.

Supplemental document. See Supplement 1 for supporting content.

References

1. L. Rahib, B. D. Smith, R. Aizenberg, *et al.*, "Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the united states," Cancer Res. **74**(11), 2913–2921 (2014).
2. S. T. Chari, "Detecting early pancreatic cancer: problems and prospects," Semin. Oncol. **34**(4), 284–294 (2007).
3. J. D. Kang, S. E. Clarke, and A. F. Costa, "Factors associated with missed and misinterpreted cases of pancreatic ductal adenocarcinoma," Eur. Radiol. **31**(4), 2422–2432 (2021).

Biomedical Optics EXPRESS

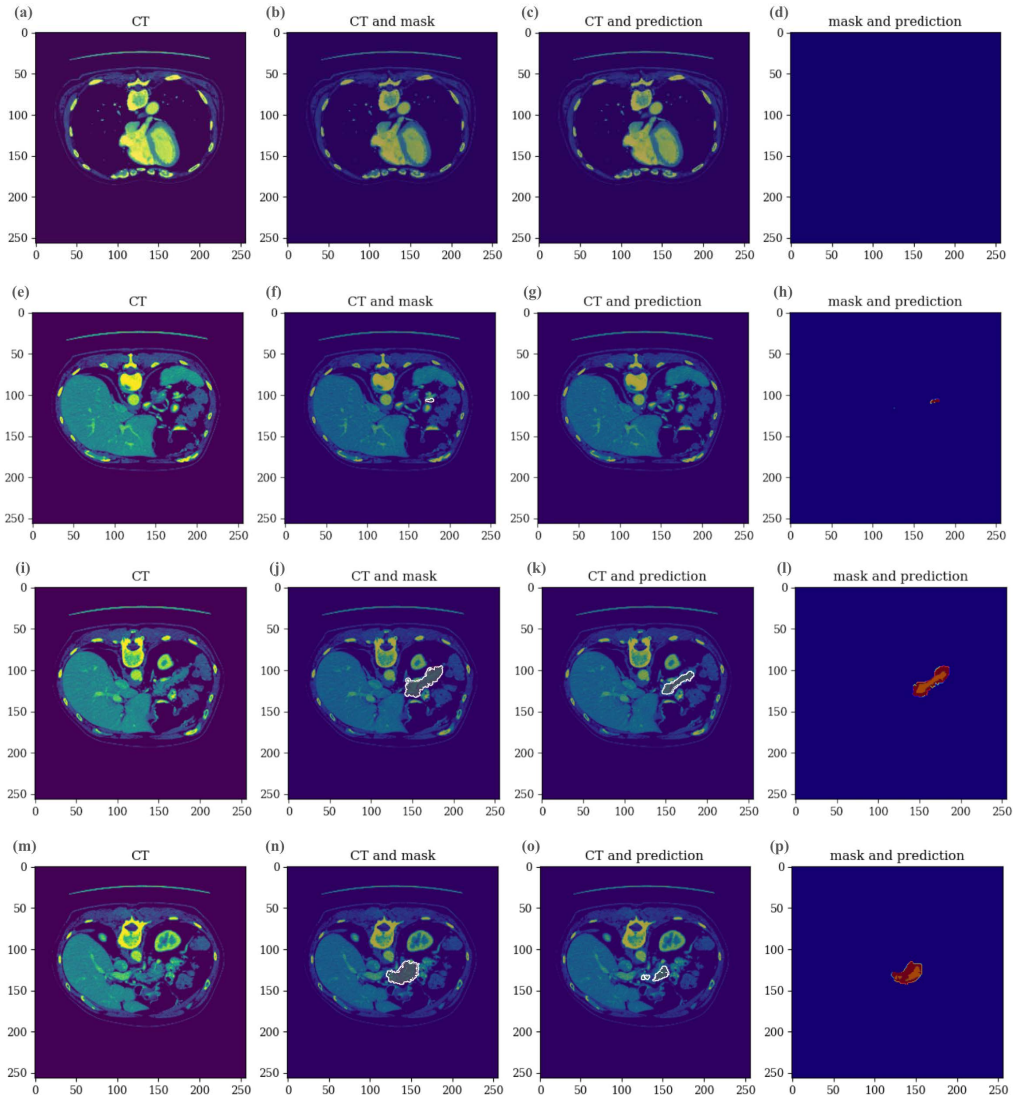


Fig. 10. Visualization of patient predictions with 3D MediONN. The figure follows the scan sequence from top to bottom, illustrating cases from the absence of tumors to their appearance. From left to right, each column represents the original CT scan (a, e, i, m), the ground truth mask overlaid on the CT scan (b, f, j, n), the predicted mask (c, g, k, o), and a comparison of both the ground truth and predicted masks (d, h, l, p).

4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Inc., 2012).
5. K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 770–778.
6. J. Redmon, S. Divvala, R. Girshick, *et al.*, "You only look once: Unified real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 779–788.
7. A. Esteva, B. Kuprel, R. A. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," Nature **542**(7639), 115–118 (2017).
8. P. Rajpurkar, J. Irvin, K. Zhu, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv (2017).

Biomedical Optics EXPRESS

9. Y.-H. Chen, T. Krishna, J. S. Emer, *et al.*, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE J. Solid-State Circuits **52**(1), 127–138 (2017).
10. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, *et al.*, "Design of ion-implanted mosfet's with very small physical dimensions," Proc. IEEE **87**(4), 668–678 (1999).
11. M. M. Waldrop, "More than moore," Nature **530**(7589), 144–147 (2016).
12. Y. Shen, N. C. Harris, S. Skirlo, *et al.*, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
13. M. Bagherian, P. Rezaei, A. Yavari, *et al.*, "Photonic neural networks: A review on hardware architectures," IEEE Access **11**, 58027–58045 (2023).
14. C. Feng, J. Gu, H. Zhu, *et al.*, "A compact butterfly-style silicon photonic–electronic neural chip for hardware-efficient deep learning," ACS Photonics **9**(12), 3906–3916 (2022).
15. Y. Zhou, W. Wang, Y. Jiang, *et al.*, "Scalable photonic computing with chip-scale architectures," IEEE J. Sel. Top. Quantum Electron. **28**, 1–14 (2022).
16. Y. Lin, Y. Wang, J. Li, *et al.*, "Chip-based training of photonic neural networks," Nat. Photonics **17**, 132–140 (2023).
17. J. Gu, Z. Zhao, C. Feng, *et al.*, "Toward hardware-efficient optical neural networks: Beyond FFT architecture via joint learnability," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **40**(9), 1796–1809 (2021).
18. B. Shi, N. Calabretta, and R. Stabile, "A nonlinear activation function for optical neural networks using a semiconductor optical amplifier–mach–zehnder interferometer," APL Photonics **7**(1), 010801 (2022).
19. H. R. Roth, L. Lu, A. Farag, *et al.*, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *MICCAI 2015, Part I*, vol. 9349 of *LNCS* N. Navab, et al., eds., (2015), pp. 556–564.
20. O. Ronneberger, P. Fischer, and A. Becker, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, (2015), pp. 234–241.
21. O. Çiçek, A. Abdulkadir, S. Lienkamp, *et al.*, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, (2016), pp. 424–432.
22. S. Chetlur, C. Woolley, P. Vandermersch, *et al.*, "Cudnn: Efficient primitives for deep learning," arXiv (2014).
23. S. Salehi, X. Zhou, X. Lin, *et al.*, "A new loss function for imbalanced classification problems," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, (2017).
24. L. N. Smith, "A better way to train neural networks: The one cycle policy," in *Proceedings of the 2018 Workshop on Machine Learning and Systems*, (MLSys, 2018), pp. 1–4.
25. X. Xu, H. Yun, Y. Wang, *et al.*, "Low-loss and broadband silicon optical directional coupler using asymmetric waveguide widths," Opt. Express **23**(3), 3795–3808 (2015).
26. A. N. Tait, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**(1), 7430 (2017).
27. N. C. Harris, Y. Ma, T. Baehr-Jones, *et al.*, "Efficient, compact and low loss thermo-optic phase shifter in silicon," Opt. Express **22**(9), 10487–10493 (2014).
28. J. Feldmann, N. Youngblood, C. D. Wright, *et al.*, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**(7755), 208–214 (2019).
29. T. Zhou, Y. Chen, C. Xu, *et al.*, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nat. Photonics **15**, 367–373 (2020).
30. S. Y. Siew, T. Y. Liow, A. E.-J. Lim, *et al.*, "Review of silicon photonics technology and platform development," J. Lightwave Technol. **39**(13), 4374–4389 (2021).
31. Q. Yan, H. Ouyang, Z. Tao, *et al.*, "Multi-wavelength optical information processing with deep reinforcement learning," Light: Sci. Appl. **14**(1), 160 (2025).
32. Z. Fan, J. Lin, T. Zhang, *et al.*, "Efficient off-chip configuration method for scalable programmable photonic integrated circuits," Commun. Phys. **8**(1), 218 (2025).