Research Article | Vol. 29, No. 20 / 27 Sep 2021 / Optics Express 31924

Optics EXPRESS

# On-chip photonic diffractive optical neural network based on a spatial domain electromagnetic propagation model

TINGZHAO FU, YUBIN ZANG, HONGHAO HUANG, ZHENMIN DU, CHENGYANG HU, MINGHUA CHEN, SIGANG YANG, AND HONGWEI CHEN*

*Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
*chenhw@tsinghua.edu.cn*

**Abstract:** An integrated physical diffractive optical neural network (DONN) is proposed based on a standard silicon-on-insulator (SOI) substrate. This DONN has compact structure and can realize the function of machine learning with whole-passive fully-optical manners. The DONN structure is designed by the spatial domain electromagnetic propagation model, and the approximate process of the neuron value mapping is optimized well to guarantee the consistence between the pre-trained neuron value and the SOI integration implementation. This model can better ensure the manufacturability and the scale of the on-chip neural network, which can be used to guide the design and manufacturing of the real chip. The performance of our DONN is numerically demonstrated on the prototypical machine learning task of prediction of coronary heart disease from the UCI Heart Disease Dataset, and accuracy comparable to the state-of-the-art is achieved.



## 1. Introduction

Deep learning is one of the fastest-growing areas in machine learning and has been extensively used in various fields. In many applications, deep neural networks (DNNs) can play an important role in addressing the issue of speech recognition [1], language translation [2], image classification [3], and medical image analysis [4]. Beyond the above applications, DNNs are also widely used to solve inverse design problems [5–11]. Of particular concern, DNNs have a wide range of applications and excellent performance, however, it requires higher computing power with the increase of the complexity of the target tasks. Therefore, it is of great significance to find alternative calculation methods that are faster and lower energy consumption. Fortunately, in recent years, the emergence of optical neural networks (ONNs) provides a way to solve the dilemma of high energy consumption when facing complex tasks. Compared with the electrical neural networks, the ONNs have the characteristics of low-power consumption, ultra-broad bandwidth, ultra-high-speed, and capability of parallel processing signals [12–24]. Previous studies have been reported, including diffractive deep neural networks ($D^2NN$) created through 3D-printed for implement classification of images of handwritten digits and fashion products based on free-space diffraction [13], task-specific accelerators based on free-space optics for massively parallel and real-time information processing [25,26], integrated photonic platforms using programmable waveguide interferometer meshes [12,15,22,23], artificial neural computing using a nanophotonic neuron medium [18], and integrated spiking neural networks using phase-change materials (PCM) [17]. However, obstacles to the development of ONNs still exist, one of which is the passive low-loss miniature integration. As to the issue of miniature integration, metasurfaces provide an exceptional solution [13,24,27,28], including one dimensional (1D) metasurfaces [24] and two dimensional (2D) metasurfaces [13,27,28]. As to the 1D metasurfaces

#435183 https://doi.org/10.1364/OE.435183
Journal © 2021 Received 24 Jun 2021; revised 25 Aug 2021; accepted 7 Sep 2021; published 21 Sep 2021

Optics EXPRESS

based on the SOI platform consisted of several metalines, each of the metaline consists of a series of slots, however, there is mutual interference between adjacent slots. As to the 2D metasurface, the misalignment of layered metasurfaces can have a devastating impact on the performance of the system as errors accumulate during propagation.

In this work, we propose an optical deep learning framework in which the neural network is physically composed of multiple layers of diffractive 1D metasurfaces, which optically perform a function that the network can statistically learn. Notably, the following two problems are considered during this research. One is the problem that the effective refractive index of the identical slot at different positions is diverse for light inputs with different angles. The second problem is the existence of mutual interference between adjacent slots of different lengths when the light inputs at the same angle. To minimize the impact of the aforementioned two problems, and make the phase delay produced by silicon slots more accurately approximate the value of the pre-trained neurons, three silicon slots are used as a single neuron, and the spacing between adjacent hidden layers (metalines) is set as 300 μm in this design. To demonstrate the capability of our DONN, we benchmark its performance on the classification of heart patients from the UCI heart disease dataset [29], which achieves an accuracy of 86.9% that is comparable to the state-of-the-art [30–35]. For verification of the results, a 2.5D variational finite difference time domain (FDTD) solver of Lumerical Mode Solution commercial software is utilized. The matching score between the simulation results and the analysis results is more than 91.8%. Our all-optical deep learning framework can promote the potential applications of photonic integrated devices in many aspects, including speech recognition, data mining, and object classification, etc.

## 2. Modeling of diffractive optical neural networks

### 2.1. Optimization of neuron value mapping

In this study, neuron value mapping refers to using the phase delay generated by slots of different sizes to represent the neuron value obtained by computer pre-training in the form of physical structure. For DONNs, it is critical to use the physical structure to accurately approximate the pre-trained neuron values. In the process of neuron value mapping, the effective refractive index (ERI) is crucial since it is adopted in the pre-training model. With such a parameter, the length of each slot required for the corresponding neuron value can be calculated according to Eq. (1):

$$L_{slot-i} = \frac{\Delta\varphi_i}{(n_{eff} - n_{slab}) \cdot k_0} \tag{1}$$

where $L_{slot-i}$ is the length of the i-th slot, $n_{eff}$ is the ERI of the slot through which light passes, $n_{slab}$ is the ERI of the slab waveguide, $k_0 = 2\pi/\lambda$ is the wave number that light travels in the slots, $\Delta\varphi_i$ is the phase delay generated by the i-th slot.

Assuming that a point light is located at the origin $(0, 0, 0)$ of the coordinates as shown in Fig. 1(a), the positive X-axis is the direction in which light propagates, the positive Y-axis is where the metaline slots are arranged. While the incident light is $x$ μm away from the metaline, the phase delay generated by the incident light passing through the same slot from different angles is diverse. Here, for convenience, the $L_{slot-i}$ is set as 2 μm, and the $n_{eff}$ can be obtained by formula $n_{eff} = \Delta\varphi_i/(L_{slot-i} \cdot k_0) + n_{slab}$ which is a variation of the Eq. (1), as depicted in Fig. 1(a). It is not difficult to find that the calculated $n_{eff}$ is different when the incident light enters the slots from diverse angles. In consequence, ensuring the incident light enters the slots at a smaller angle is conducive to maintain a relatively stable ERI ($n_{eff} = n_{eff0} = 2.166$), and is of advantage to the neuron value mapping. Increasing the distance between the input light and the metaline can reduce the incident angle of the input light relative to each slot, thus reducing the difference of the ERI calculated by the phase delay and making the approximation of the neuron value mapping more accurate.

Optics EXPRESS

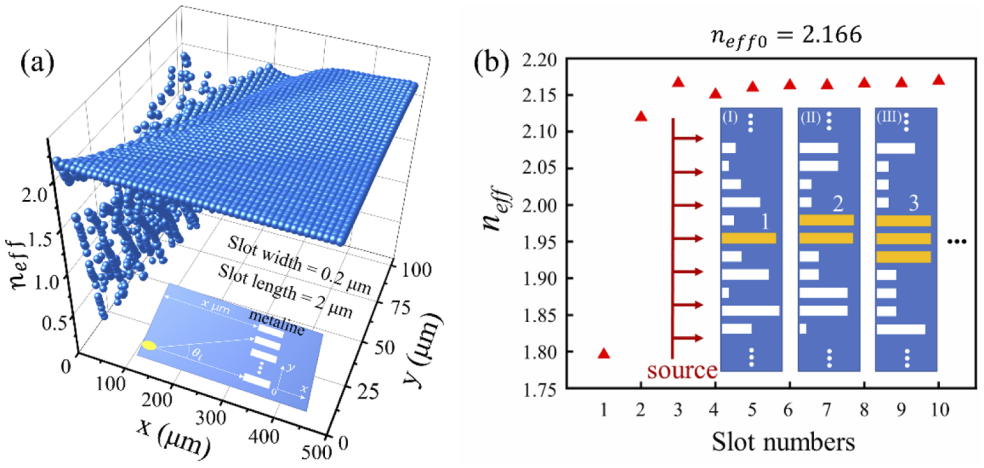


**Fig. 1.** (a) After the input light propagating $x$ μm to the metaline, the effective refractive index is calculated by the phase delay generated when the light travels through the identical slots at different angles. Of which the distance between the center of adjacent slots is 500 nm, the width of the slots is 200 nm, and the length is 2 μm. (b) Calculation of effective refractive index under a different number of identical slots in a slot group. The center distance of the adjacent slots and the width of the slots are the same as (a), while the length of the metaline slots is diverse. $\theta_i$ is the angle between the i-th incident ray and the horizontal line.

Additionally, because of the mutual interference between adjacent slots, even when the incident light passes through the slots at the same angle (for example, plane waves), the phase delay generated by the slots is different if the length of the slots is distinct, therefore the $n_{eff}$ calculated by the phase delay will also be different. In order to reduce the influence of mutual interference between adjacent slots on the process of neuron value mapping, a slot group composed of multiple identical slots is used to approximate a neuron value. The $n_{eff}$ calculated by the phase delay generated by the slot group with a different number of slots is shown in Fig. 1(b). It is easy to find that when the number of orange-yellow slots increases, the ERI (the red triangle) calculated by the phase delay generated by the slot groups tends to be a stable value ($n_{eff0}$ = 2.166). This is beneficial to the mapping process of neuron values.

In Fig. 1(b), the width of the slot in inset (I) is 200 nm, the period is 500 nm, and the length of the slot is randomly generated. Among them, the length of the slot marked in orange-yellow is 1.964 μm, and the ERI obtained by calculating the phase delay generated when the flat light passes through the slot is 1.796, which is significantly different from the ERI ($n_{eff0}$ = 2.166) used in the pre-training model. Thus, a slot group composed of multiple identical slots is proposed as shown in insets (II) and (III), the ERI obtained by calculating the phase delay generated when the flat light passes through them is 2.119 and 2.169, respectively. When the slot group includes three identical slots, the calculated ERI is 2.169, which is very close to 2.166. Moreover, in Fig. 1(b), it is not difficult to see that when the number of slots in the slot group is more than 3, the calculated ERI (the red triangle) is close to 2.166 and tends to be stable. Therefore, the process of neuron value mapping is facilitated by the use of a slot group consisting of multiple identical slots.

As mentioned above, the more stable ERI can be obtained by increasing the distance between the input light and the metaline. Here, the detailed simulation examples are shown in Fig. 2.

In Fig. 2(a) and (b), it is easy to find that the ERI of each slot calculated by FDTD is very close to the $n_{eff0}$, and it shows that the calculated ERI is insensitive to the propagation distance when the incident light travels horizontally into the slots. In Fig. 2(c), when the distance between the

Optics EXPRESS

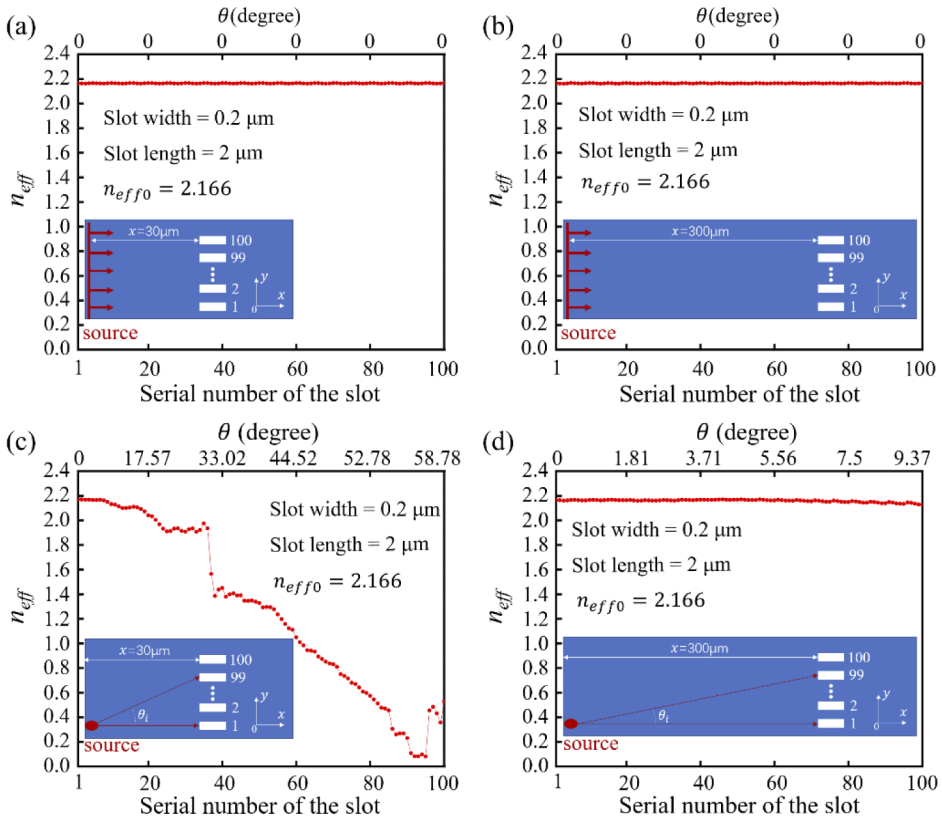


**Fig. 2.** (a) The incident light source in (a) and (b) is a flat light, and the distance between the light source and the metaline is 30 μm and 300 μm, respectively. The incident light source in (c) and (d) is a point light, and the distance between the light source and the metaline is also 30 μm and 300 μm, respectively. Among them, the metaline consists of 100 identical slots, numbered from 1 to 100. The red dots are the ERI values calculated by the phase delay $\Delta\varphi_i$ generated by the light passing through the slot of the corresponding sequence number. $n_{eff0}$ is the ERI in the pre-training model, the period of the slots is 0.5 μm, and $\theta_i$ is the angle between the i-th incident ray and the horizontal line.

light source and the metaline is 30 μm, the incident light travels into the slots at different angles ranging from 0 to 58.78°, causing the calculated ERI to be diverse. However, in Fig. 2(d), when the distance between the light source and the metaline becomes 300 μm, the incident light travels into the slots at angles ranging from 0 to 9.37°, the calculated ERI tends to be stable near $n_{eff0}$. This indicates that when the light travels into the slot at a small angle (here less than 9.5°), the calculated ERI at the slot is closer to the ERI adopted in the pre-training model, and the neuron value mapping process is optimal. Therefore, by increasing the distance between the incident light and the metaline, the more stable the ERI when the light passes through the slot (that is, the closer it is to the ERI in the pre-training model), and the more accurate the neuron value mapping can be completed.

Furthermore, a concrete simulation example to illustrate the optimization result is shown in Fig. 3. Suppose that there are three pre-trained neuron values equaling $-\pi/4$, $-3\pi/4$, $-5\pi/4$. Those values can be achieved through using the phase delays generated by the slots. From Fig. 8(a), when the width, height, and period of the selected slots are 0.2 μm, 0.22 μm, and 0.5 μm, respectively, the ERI of the slots $n_{eff} = 2.166$. Therefore, according to Eq. (1), the length

Optics EXPRESS

of the slots to realize the three pre-trained neuron values is 0.287 μm, 0.862 μm, and 1.437 μm, respectively.

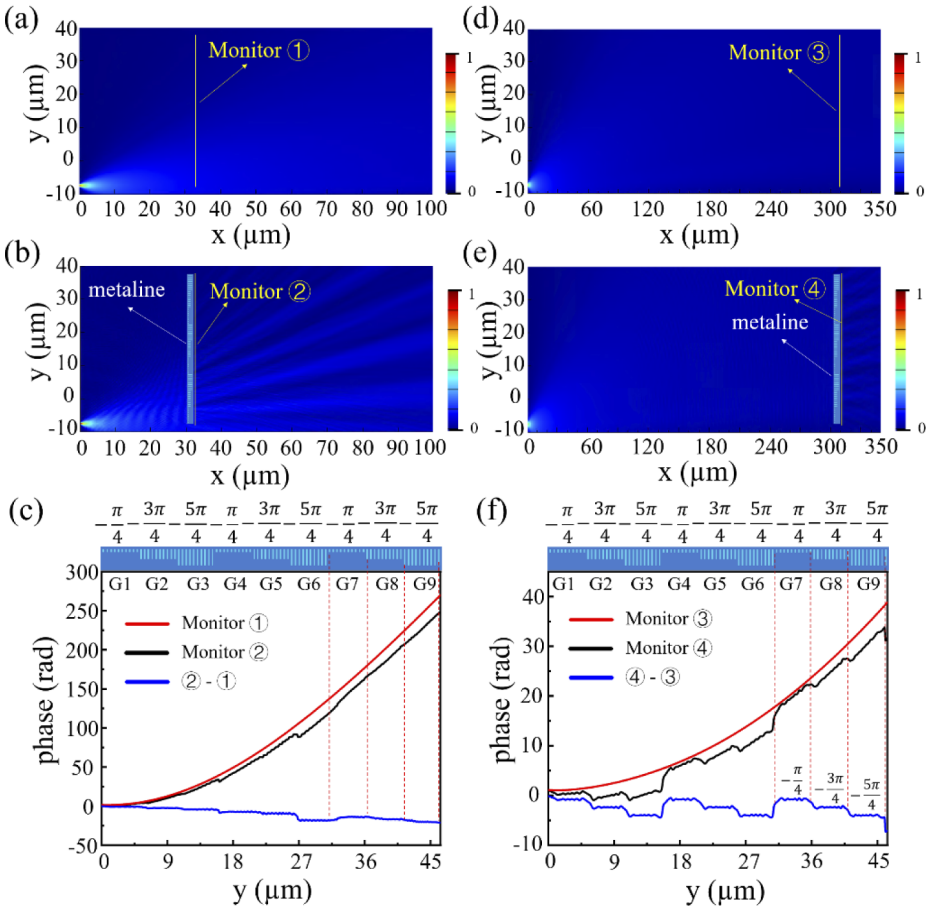


**Fig. 3.** The simulation examples to illustrate the optimization process. In Fig. 3(a) and (d), the distance from the point light source to Monitor ① and ③ is 32 μm and 302 μm, respectively. In Fig. 3(b) and (e), the same metaline is designed at 30 μm and 300 μm away from the point light source, respectively, and the monitors are located in the same position as in (a) and (d). Figure 3(c) is the numerical analysis of the monitors in Fig. 3(a) and (b). Figure 3(f) is the numerical analysis of the monitors in Fig. 3(d) and (e). Each slot group from G1 to G9 consists of 10 identical slots.

In Fig. 3(c), the distance between the light source and the metaline is 30 μm, the blue line "②-①" shows the subtraction results of the Monitor ① and ②, and the values differ greatly from those of pre-trained neurons. In the contrast, when the distance between the incident light source and the metaline becomes 300 μm in Fig. 3(f), the subtraction values of the Monitor ④ and ③ as shown by the blue line "④-③" are highly consistent with the pre-trained neurons.

By reason of the foregoing, our optimization method is effective, that is, by increasing the distance between the input light and the metaline and using a slot group composed of multiple identical slots can obtain a more stable ERI which is closer to the ERI adopted in the pre-training model and further make the approximation of the neuron value mapping more accurate.

Optics EXPRESS

Without loss of generality, the research rule of the distance between the incident light and the metaline is also applicable to the distance design between the adjacent hidden layers (metalines) in the neural network system.

### 2.2. Spatial domain electromagnetic propagation model

The spatial domain electromagnetic propagation model (SDEPM) of our DONN is adopted based on the Huygens-Fresnel principle. In this work, the thickness of the slab waveguide is only 220 nm, therefore the classical Huygens-Fresnel principle is essential to be modified according to the restricted propagation conditions. The modified SDEPM is described as shown in Eq. (2):

$$w_i^m = \frac{1}{j\lambda} \cdot \left( \frac{x - x_i}{r_i^2} \right) \cdot \exp\left( \frac{2\pi r n_{slab}}{\lambda} \right) \cdot \beta \exp(j\Delta\phi) \tag{2}$$

where $m$ represents the $m$-th layer of the network, $i$ represents the $i$-th neuron located at $(x_i, y_i)$ of layer $m$, $\lambda$ is the working wavelength, $j = \sqrt{-1}$ is an imaginary unit, $r_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$ is the distance between a neuron at layer $m$ and each neuron at layer $m - 1$, $n_{slab}$ is the ERI of the slab waveguide, $\beta$ and $\Delta\phi$ are the correction factors for the classical Huygens-Fresnel principle when light propagates a certain distance in a slab waveguide, where $\beta$ is a specific coefficient and $\Delta\phi$ is a fixed phase delay. Figure 4 shows the amplitude and phase distribution of the input signal waveform after 300 μm propagation in the slab waveguide (thickness is 220 nm, width is 105 μm) based on the modified SDEPM and FDTD, respectively. Apparently, Fig. 4(c) and (d) indicate that the electric field evolution of the input signal (Fig. 4(a) and (b)) propagating 300 μm later in the modified SDEPM is highly consistent with the simulation results of FDTD.

### 2.3. Forward and backward propagation

Following the SDEPM diffraction Eq. (2), one can consider every single neuron of a given DONN layer as a secondary source of a wave, the amplitude and relative phase of this secondary wave are determined by the product of the input wave to the neuron and its transmission coefficient ($T$). Therefore, for the $m$-th layer of the network, one can describe the output function ($n_i^m$) of the $i$-th neuron located at $(x_i, y_i)$ as:

$$n_i^m(x_i, y_i) = w_i^m(x_i, y_i) \cdot T_i^m(x_i, y_i) \cdot \sum\nolimits_k n_k^{m-1}(x_i, y_i) \tag{3}$$

where $\sum\limits_k n_k^{m-1}(x_i, y_i)$ is the input wave to the $i$-th neuron of layer $m$, $T_i^m(x_i, y_i)$ is the transmission coefficient of $i$-th neuron of layer $m$. Here $T_i^m(x_i, y_i)$ can be described as:

$$T_i^m(x_i, y_i) = a_i^m(x_i, y_i) \cdot \exp(j\varphi(x_i, y_i)) \tag{4}$$

where $a_i^m(x_i, y_i)$ is the amplitude of the transmission coefficient, in this study, $a_i^m$ is set as the constant 1 because the optical loss of each neuron of layer $m$ is negligible, $\varphi(x_i, y_i)$ is the phase factor of the corresponding neuron.

The forward propagation model can be obtained as shown in the Eq. (5):

$$\begin{cases} n_{i,p}^m = w_{i,p}^m \cdot T_i^m \cdot u_i^m \\ u_i^m = \sum_k n_{k,i}^{m-1} \\ T_i^m = a_i^m \cdot \exp(j\varphi_i^m) \end{cases} \tag{5}$$

where $i$ represents a neuron of the $m$-th layer, and $p$ refers a neuron of the next layer, connected to neuron $i$ by optical diffraction. Excluding the input and output layers, assuming that the DONN

Optics EXPRESS

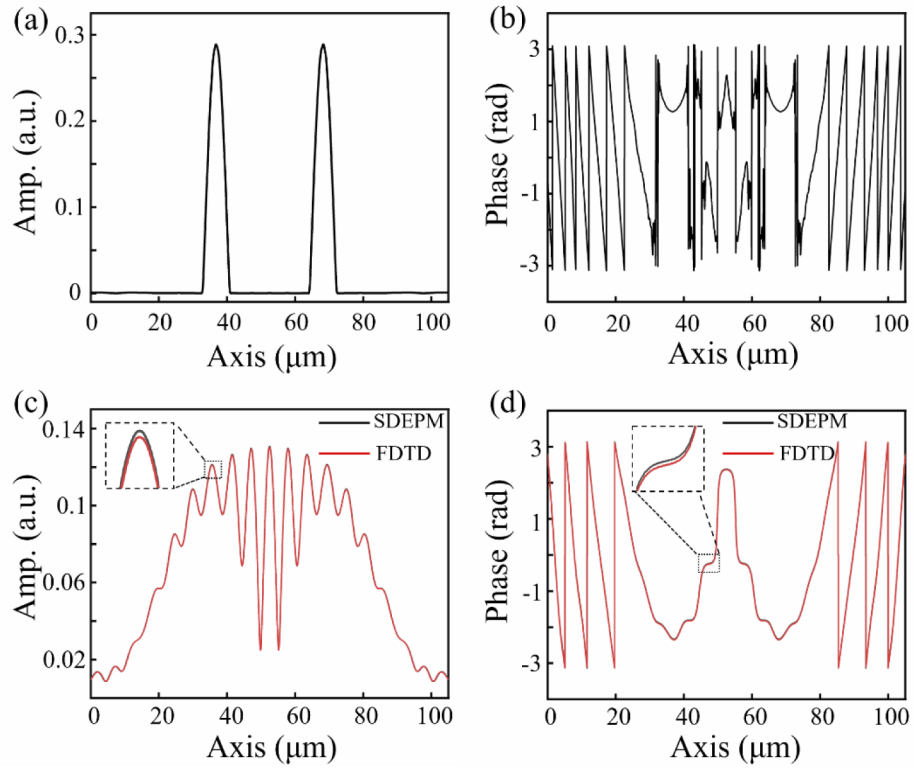


**Fig. 4.** (a) and (b) are the field intensity and phase distribution of the input signal, respectively. (c) and (d) are the field intensity and phase distribution of the input signal propagating 300 μm later in a slab waveguide of the modified SDEPM (black line) and 2.5D variational FDTD (red line), respectively.

consists of $M$ layers, then the electric field of the $p$-th neuron in the output layer can be obtained by Eq. (6) and the corresponding intensity of the resulting optical field can be measured by a detector as Eq. (7):

$$u_i^{M+1} = \sum\nolimits_k n_{k,i}^M \tag{6}$$

$$S_i^{M+1} = \left|u_i^{M+1}\right|^2 \tag{7}$$

Once the intensity of the resulting optical field is computed, the loss function in our optimization can be defined as the normalized mean square error (NMSE) to measure the distance which can be described as between the desired target intensity $Tar_i^{M+1}$ and the realized test intensity $S_i^{M+1}$ of the output areas. Here, we define the loss function ($L$) as:

$$L(\varphi_i^m) = \frac{1}{N}\sum\nolimits_k \left(\frac{S_k^{M+1}}{\sum_k S_k^{M+1}} - Tar_k^{M+1}\right)^2 \tag{8}$$

where $N$ refers to the number of measurement points at the output plane. Consequently, the problem of the optimization for a DONN design can be summarized as $\min L(\varphi_i^m)$, where the range of $\varphi_i^m$ is over 0 to $2\pi$. Then, the back propagation algorithm is used to train the phase value at each hidden layer in the DONN network. The gradient calculation formula of the loss function

Optics EXPRESS

is shown as follows:

$$\frac{\partial L}{\partial \varphi_i^m} = \frac{4}{N}\sum_k \left(\frac{S_k^{M+1}}{\sum_k S_k^{M+1}} - Tar_k^{M+1}\right) \cdot \frac{\sum_k S_k^{M+1} - S_k^{M+1}}{\left(\sum_k S_k^{M+1}\right)^2} \cdot Real\left\{(u_k^{M+1})^* \cdot \frac{\partial u_k^{M+1}}{\partial \varphi_i^m}\right\} \quad (9)$$

where $u_k^{M+1}$ quantifies the gradient of the complex-valued optical field at the output layer with respect to the phase values of the neuron in the previous layers, $m \le M$. In Eq. (9), $\frac{\partial u_k^{M+1}}{\partial \varphi_i^m}$ can be calculated as follows:

$$\frac{\partial u_k^{M+1}}{\partial \varphi_i^{m=M-L}} = j \cdot T_i^{M-L} \cdot u_i^{M-L} \cdot \sum_{k_1} w_{k_1,k}^M \cdot T_{k_1}^M \cdot \cdots\cdots \sum_{k_L} w_{k_L,k_{L-1}}^{M-L+1} \cdot T_{k_L}^{M-L+1} \cdot w_{i,k_L}^{M-L} \quad (10)$$

where, $3 \le L \le M-1$. So much for, the neuron parameters on each hidden layer can be trained and obtained by the error backpropagation algorithm. Without loss of generality, the flow chart of the forward propagation and error back propagation is depicted in Fig. 5.

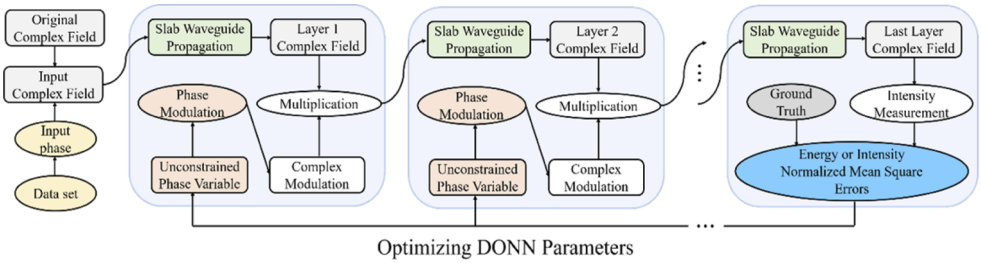


**Fig. 5.** The attributes of the data set are mapped to the phase (0 *to* $\pi$) and then the mapped phase is modulated to the original complex field to obtain a new input complex field. The resulting complex field of slab waveguide propagated field is multiplied with a complex modulator at each layer and is then transferred to the next layer. To help with the DONN design, a normalized mean square errors function is used to measure the distance which can be described as between the desired target intensity and the realized test intensity of the output.

## 3. Architecture design of diffractive optical neural networks

A conventional artificial neural network consists of an input layer, hidden layer(s), and output layer. The input and output layer generally includes one or more inputs or outputs, while each hidden layer contains lots of neurons, as shown in Fig. 6(a). For ONNs, the difference is that the hidden layers are composed of many units (or meta-atoms) of transmission or reflection, the values of each neuron on different hidden layers are often set and updated by changing the amplitude, phase, polarization, and other factors of the meta-atoms. The schematic of ONNs is depicted in Fig. 6(b).

In this study, the DONN architecture consists of one or more layers of metalines, its physical network is all-optical, and can be realized to solve complex tasks through the interference of transmitted light. The physical structural parameters that implement the interference and prediction mechanisms are designed in advance. Firstly, the parameters of the neural network structure are trained on the computer and then these parameters are mapped to the physical structure of the DONN. A schematic view of our proposed DONN design is presented in Fig. 7. Here, a single neuron is formed by three identical slots, which is named meta-atom. The thickness of the silicon ($Si$) substrate $t_1$ is 3 μm, the silicon dioxide ($SiO_2$) insulator layer $t_2$ is 2 μm, and

Optics EXPRESS

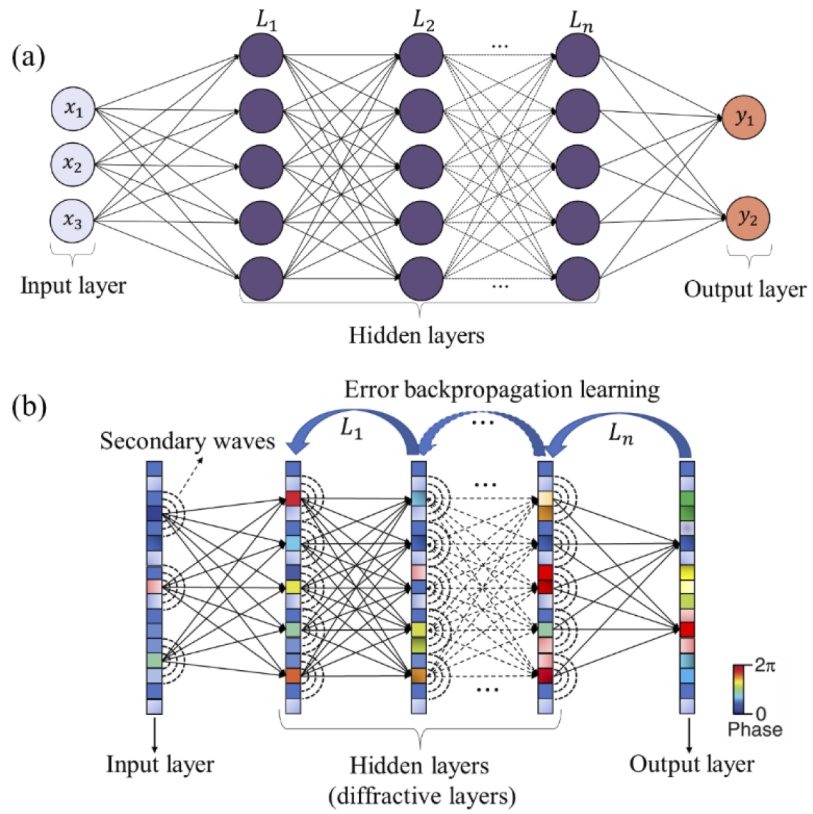


**Fig. 6.** (a) Schematic of a conventional artificial neural network. (b) Schematic of diffractive optical neural network, each point on a given layer acting as a secondary source of a wave, the amplitude and phase of which are determined by the product of the input wave and the complex-valued transmission or reflection coefficient at that point.

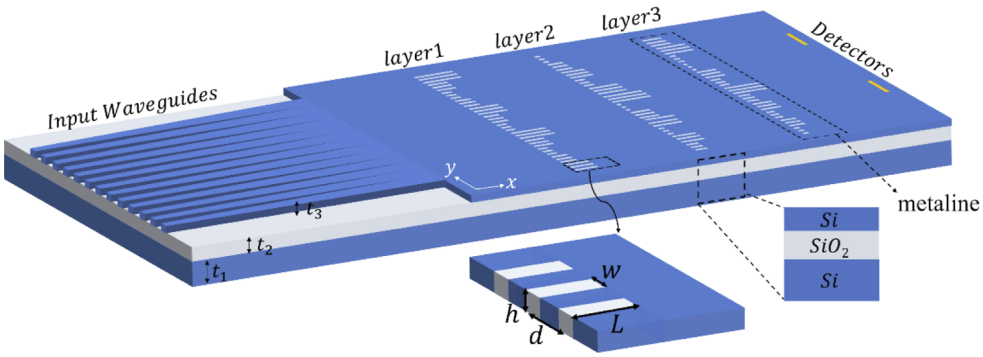


**Fig. 7.** Schematic of the diffractive optical neural network. A DONN constitutes three hidden layers (metalines), where each of the three slots on a given layer act as a neuron, with a complex-valued transmission coefficient. The transmission coefficients of each layer can be trained by using deep learning to perform a function between the input and output plane of the network. After this learning phase, the DONN design is fixed; once manufactured by electron beam lithography and other microelectronic processes, it performs the learned function at the speed of light.

Optics EXPRESS

the *Si* waveguide layer $t_3$ is 220 nm. In addition, there is a 2 μm layer of $SiO_2$ on the entire *Si* waveguide layer, which is not shown in the schematic.

In Fig. 7, the lattice constant $d$ of the metalines is fixed to be less than half of the wavelength, which is chosen as 500 nm. Theoretically, specific phase delays can be achieved by the corresponding slots with different lengths, widths, and heights, due to the ERI of the slot varies depending on the size [24,36]. In this work, the operation wavelength is 1.55 μm, the effective media theory (EMT) [36–38] and Lumerical FDTD are respectively utilized to approximate the ERI of the periodic slot array, and the ERI changes under the different slot widths are shown in Fig. 8(a). Next, by fixing the width of the slots to 200 nm, the free control of the propagation phase is achieved within the range from 0 to 2π by changing the length of the slots from 0 to 2.3 μm. Furthermore, the transmission amplitude is higher than 94% obtained by the commercial software package Lumerical FDTD under the corresponding length of the slot, as shown in Fig. 8(b).

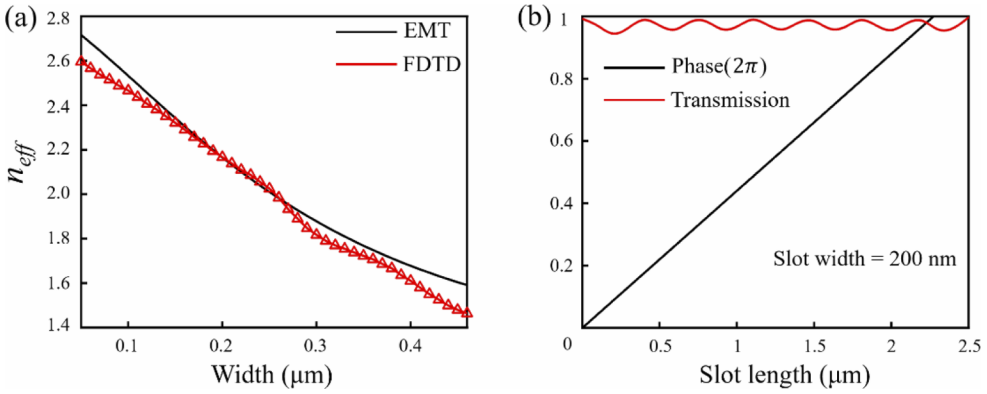


**Fig. 8.** (a) The effective refractive index of the different slot arrays under various slot width. Black curve is the EMT calculation; Red curve is the Lumerical FDTD simulation. (b) The simulated phase and amplitude changes versus different slot length of the slot arrays, fixing the slot width to be 200 nm.

In this study, to demonstrate the capability of our DONN, five neuron network structures are designed to verify the classification of coronary heart disease (CHD) from the UCI heart disease dataset. Firstly, the input eigenvalues are modulated onto the phase of the input light, and then the new dataset with phase information is utilized to train the parameter values of each neuron on each hidden layer through the adaptive moment estimation (Adam) optimizer. Next, the values of neuron parameters pre-trained are mapped onto silicon-based slots with different phase delays, and to avoid the mutual interference of the adjacent slots as much as possible, three identical silicon slots are used to approximate a single neuron value in this design. In addition, to make the approximating process of the neuron value mapping more accurate, the distance between the layers (input layer, hidden layers, and output layer) is also taken into account.

The optimized DONNs consist of 1 to 5 metaline layers, respectively, with each metaline (105 μm length) containing 70 neurons (consists of 210 slots). The distance between two successive metalines is 300 μm. The input signal will be loaded onto the corresponding input waveguides and propagated 160 μm through the Taper into the slab waveguide, then 300 μm through the slab waveguide to reach the first hidden layer. After light exits the final metaline (the last hidden layer), it also propagates 300 μm until it reaches the output layer of the network with two detector regions (“D1” and “D2”) arranged in a linear configuration. A specific category is assigned to each detector. The width of each detector is 8 μm, and the distance between the center of the two neighboring detectors is 40 μm. For each category, the desired intensity distribution is defined at

Optics EXPRESS

the output layer of the DONNs as a door function distribution (including forty points, twenty successive "1" or "0" in a row).

## 4. Verification of the designed DONNs and discussion

### 4.1. Numerical calculation and simulation results

To numerically demonstrate the performance of our DONNs, the prototypical machine learning task of prediction of CHD from the UCI heart disease dataset is utilized. The dataset consists of 303 sets of data, each of which contains 13 input attributes and 1 output attribute ("1" or "0"), meanwhile, "1" represents "Patient", and "0" represents "Normal" [29]. Here, the dataset is divided into training set and test set according to the rule of 8:2, that is, the training set has 242 sets of data and the test set has 61 sets of data. Thirteen input eigenvalues will be loaded onto the corresponding input narrow waveguides in the form of phases, and the predicted results are two categories, including "Normal" and "Patient" (denoised as "0" and "1", respectively). The parameters in the whole system of DONNs are trained in advance. Once the design is finalized and manufactured, the working process of the DONNs is fully optical.

Five DONNs are optimized, with each included one, two, three, four and five hidden layers, respectively. Here, we denote the neural network system with m hidden layers as DONN-m (m = 1,2,3,4,5). Figure 9 shows the loss values for the training set and the accuracy values for the test set during the learning procedure. The phase profile of each hidden layer of DONN-1, DONN-2 and DONN-3 are depicted as Fig. 10, respectively. Each hidden layer of the DONN includes 70 phase elements ranging between 0 and $2\pi$, here the linear phase profile for each layer is reshaped to an image with $7 \times 10$ pixels.

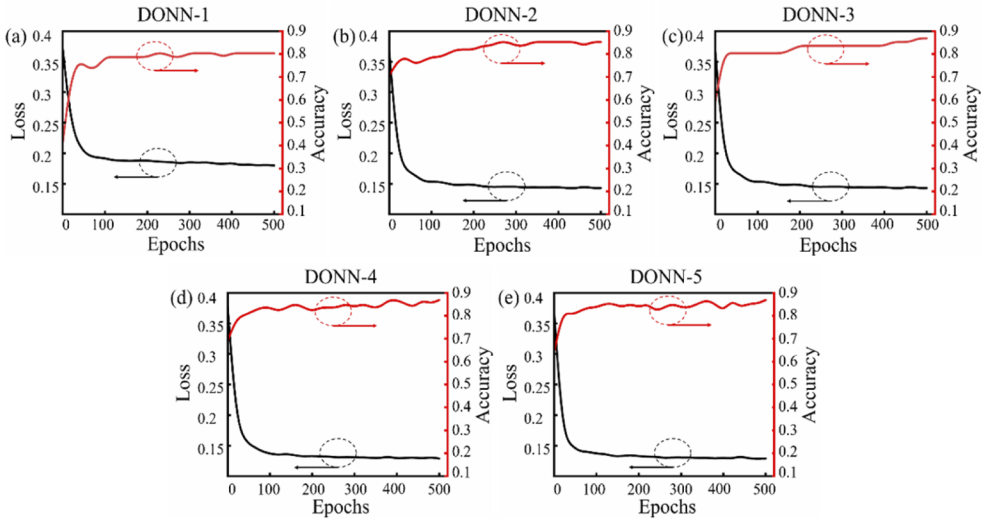


**Fig. 9.** The loss curves on the training set (black line) and accuracy curves on the test set (red line) for the optimized DONN-m during the learning procedure, while DONN-m indicates that the diffractive optical neural network system contains m hidden layers.

Furthermore, as depicted in Fig. 11, it is easy to find that when the number of hidden layers increases to more than 2, the accuracy of the test set no longer improves too much. Therefore, in terms of classification accuracy and power consumption, based on the minimum requirements, two or three hidden layers of the optimal design of the DONN may be feasible and sensible.

To illustrate the overall performance of our DONN design, 2.5D variational FDTD is used to simulate and verify the performance of DONN-1, DONN-2 and DONN-3. Figure 12 shows the

Optics EXPRESS

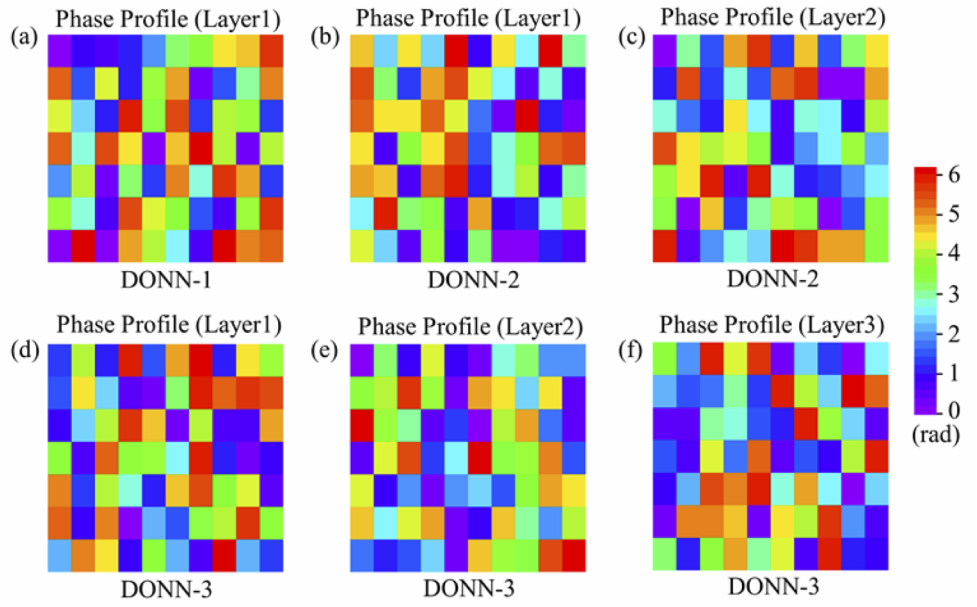


**Fig. 10.** The linear phase profile of each hidden layers is converted to a $7 \times 10$ pixelated image. (a) is the phase profile of DONN-1 hidden layer after training. (b) and (c) are the phase profiles of DONN-2 hidden layers after training. (d)~(f) are the phase profiles of DONN-3 hidden layers after training.

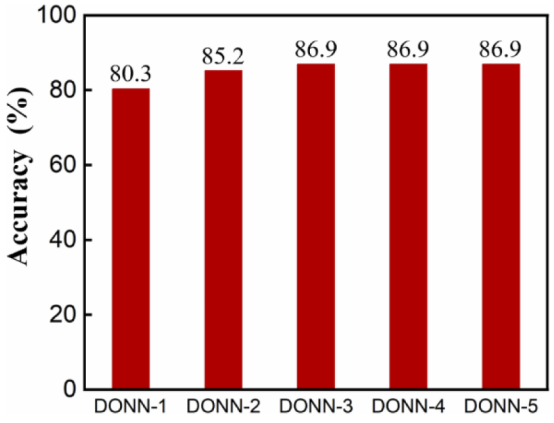


**Fig. 11.** The accuracy (test set) of the proposed DONN-m architecture on the UCI heart disease dataset. DONN-m means that the DONN architecture contains m hidden layers, here m = 1, 2, 3, 4, 5, respectively.

Optics EXPRESS

light field propagation in a two-layer hidden layer before and after training. It can be seen that before the training, light is directed to a random distribution. After training, light is focused on the right classification area. Figure 13 shows the 2.5D variational FDTD simulation results of the DONN-1, DONN-2 and DONN-3, respectively.

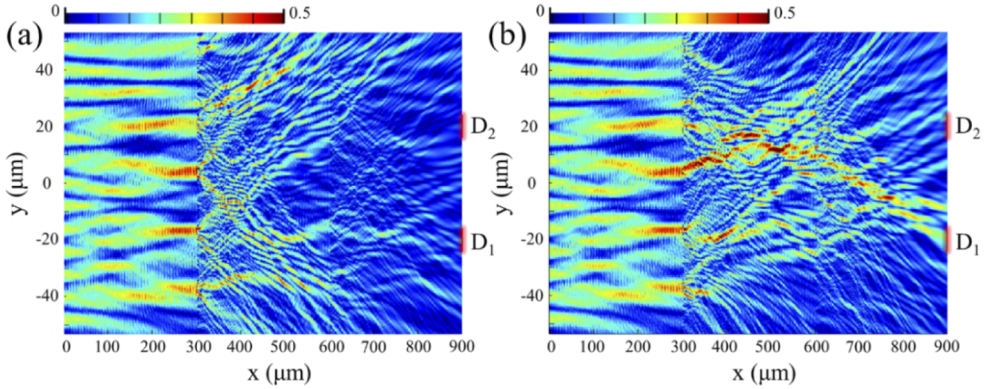


**Fig. 12.** Comparison of light propagation (a) before and (b) after training for the two-layer hidden layer. The input signal corresponds to the predicted result of the "Patient" state, which means that the detection light intensity of D1 should be greater than that of D2. (a) Before training, the lengths of the slots are randomly initialized, and the transmitted light is randomly distributed at the output plane. (b) After training, the transmitted light is directed to D1, the power of D1 is greater than that of D2, which corresponds to the input signal, and the prediction result is right.

The confusion matrices for the chosen 61 samples from the UCI heart disease dataset (test set) of the DONN-1, DONN-2, and DONN-3 are depicted in Fig. 14. Comparison between the prediction results based on the SDEPM and the simulation results based on the 2.5D variational FDTD is shown in Table 1.

**Table 1. Comparison of prediction results between SDEPM and 2.5D variational FDTD**

| DONNs | Test dataset accuracy of the SDEPM | Test dataset accuracy of the 2.5D variational FDTD | Matching score |
|---|---|---|---|
| DONN-1 | 80.3% | 78.7% | 95.1% |
| DONN-2 | 85.2% | 82% | 93.4% |
| DONN-3 | 86.9% | 83.6% | 91.8% |

### *4.2. Discussion*

In the simulation, to avoid the mutual interference caused by adjacent slots, three identical slots are utilized to approximate a pre-trained neuron value. However, the simulation results indicate that the test dataset accuracy of the 2.5D variational FDTD is lower than that of the SDEPM, obviously, the approximation of a single neuron value represented by three slots inevitably exists irreversible errors. It is easy to find that as the number of hidden layers increases, the matching score of the two results decreases, and the error caused by the approximation will be larger. Additionally, according to the numerical calculation based on the SDEPM, with the increase in the number of hidden layers, the test dataset accuracy increases very little, thus when designing the neural network structure, various factors including test dataset accuracy, accumulation of approximate errors, and power consumption, etc. should be taken into consideration.

Optics EXPRESS

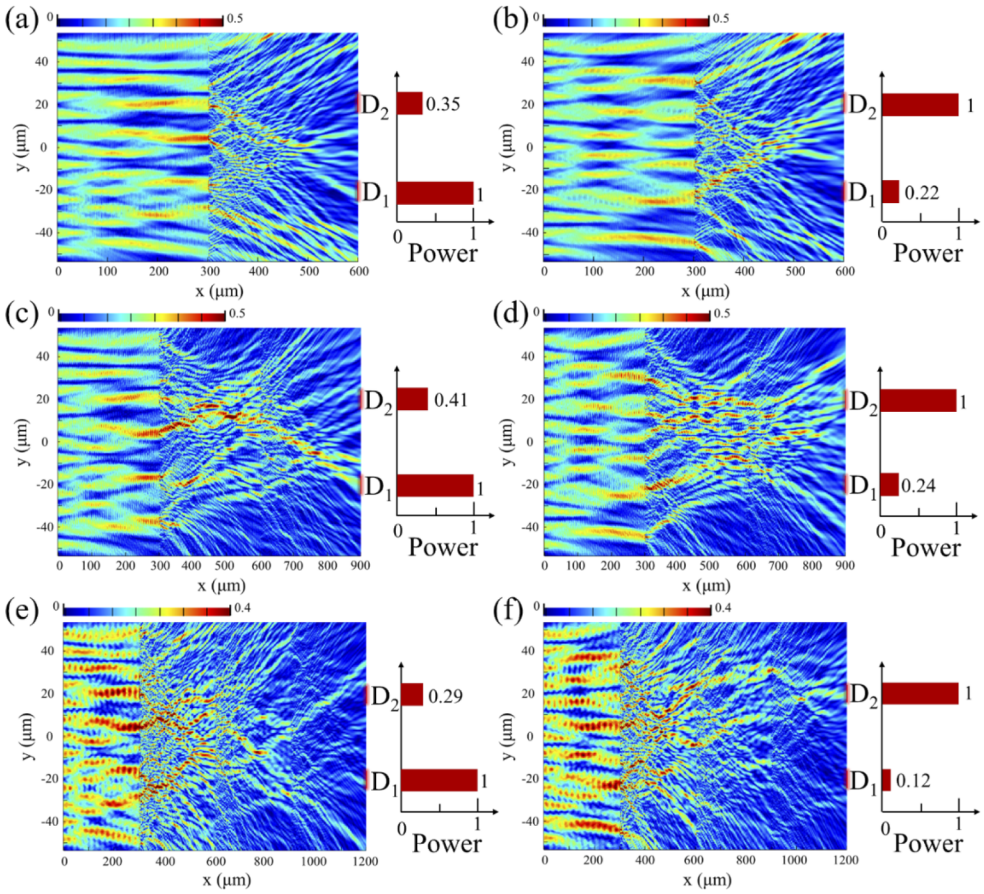


**Fig. 13.** The simulation results of the proposed DONN-1, DONN-2, and DONN-3. (a), (c), and (e) are the predictions of "Patient" state, in other words, the power of D1 should be greater than that of D2, the prediction results are right; (b), (d), and (f) are the predictions of "Normal" state, the power of D1 should be less than that of D2, the prediction results are right.

Once the optimized DONN is designed and manufactured based on an SOI substrate, it is fully optical and can perform computations on the optical signals without additional energy input (except for the energy required to input a signal). Therefore, the power consumption of the fixed DONN is only determined by the optical source which supplies the input signal (input source, $E^{in}$). For the loss of the DONN, it mainly includes the propagation loss and transmission loss of the metalines. In our design, based on Lumerical Mode Solution simulations, the loss per metaline (hidden layer) is about 0.2 $dB$.

For the latency of the DONN, it is defined as the overall time between the start of the loading signal (input source, $E^{in}$) and the detection of the output signal (computing an inference result, $E^{out}$). In a nutshell, the defined latency is the travel time for an optical input through all layers. In our DONN, the latency can be calculated by the Eq. (11) [24]:

$$Latency = D_{nw} \times c_1^{-1} + M \times D_m \times c_2^{-1} + (D_{wf} + (M-1) \times D_p + D_f) \times c_3^{-1} \tag{11}$$

where $D_{nw}$ is the distance from the narrow waveguide to which the signal is loaded to the slab waveguide; $c_1 = \frac{c_0}{n_{eff1}}$ is the speed of light in the silicon narrow waveguide (thickness is 220 nm,

Optics EXPRESS

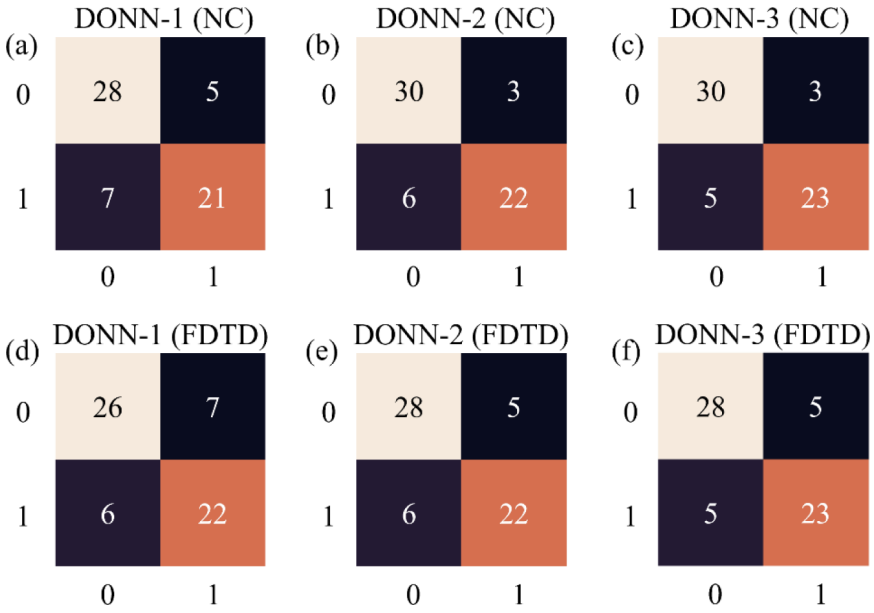


**Fig. 14.** (a) Confusion matrixes for the chosen 61 samples from the heart disease dataset (test set) of the DONN-1 (NC), DONN-2 (NC) and DONN-3 (NC) generated based on the results of SDEPM, meanwhile, "NC" means the numerical calculation; (b) Confusion matrixes of the DONN-1 (FDTD), DONN-2 (FDTD) and DONN-3 (FDTD) generated based on the results of 2.5D variational Lumerical Mode Solution simulations.

width is 450 nm), where $c_0$ is the vacuum light speed, $n_{eff1}$ is the effective index of refraction of the narrow waveguide; $M$ is the total number of the hidden layers; $D_m$ is the maximum propagation distance through each layer; $c_2 = \frac{c_0}{n_{eff2}}$ is the speed of light in the metalines, where $n_{eff2}$ is the effective index of the silicon slots; $D_{wf}$ is the distance between the interface of the narrow waveguide and the slab waveguide to the first hidden layer; $D_p$ is the propagation distance between the hidden layers; $D_f$ is the propagation distance between the last hidden layer and the output layer; $c_3 = \frac{c_0}{n_{eff3}}$ is the speed of light in the slab waveguide (the silicon device layer), where $n_{eff3}$ is the effective index of the slab waveguide. As an example, for our designed DONN-3, $n_{eff1} = 2.33$, $n_{eff2} = 2.166$, $n_{eff3} = 2.84$, the latency is approximately 12.65ps.

For the computational speed of the DONN, it is defined as the number of input vectors that can be processed per unit time. Crucially, the speed of our DONN structure is limited by the photodetectors. By assuming that the rate of the photodetector is 25GHz, the DONN with $N$ neurons per layer can perform $N^2 \times L \times 2.5 \times 10^{10}$ MAC/sec. In our study, each hidden layer contains 70 neurons ($N = 70$), as an example, for our designed DONN-3 ($L = 2$), the computational rate is about $2.45 \times 10^{14}$ MAC/sec. This is two orders of magnitude higher than the performance of modern GPUs, which typically perform $10^{12}$ floating-point operations per second [22].

In our design, the width of the DONNs is 105 μm, and the footprint of the DONN-m is about $105\ \mu m \times (160\ + (m+1) \times 300\ )\ \mu m$ ($m = 1, 2, 3, \ldots$). For the scale of trainable parameters, we propose DONN with 1-3 metaline layers, each containing 210 slots. Since three slots are regarded as one neuron whose value is trainable in order to suppress interference, there are 70 trainable parameters in each metaline layer. The proposed DONN-3, which includes three metalines composed of 210 trainable neurons (630 slots), has a CHD classification accuracy of 86.9%. This accuracy is comparable to the state-of-the-art (80%~89%) [30–35]. In this study,

Optics EXPRESS

non-linear activation is not used, thus, it may further improve the performance of the system if nonlinear activation function is used.

Finally, despite our DONNs architecture have many advantages, such as passive miniature integration, without alignment between hidden layers, simple structure, and ease of massive manufacturing capability, lower power consumption and light-speed processing, etc., the number of parallel channels of input signals is limited, and the accumulation of approximate errors has a greater impact on performance as the number of hidden layers increases. It is still necessary to search for new basic physical structural units to conduct better approximation of the pre-trained neuron values.

## 5. Conclusion

In summary, we put forward a whole-passive fully-optical DONN architecture based on SOI, the pre-trained neuron values are mapped onto the different phase delays in terms of physical structures, and the corresponding phase delays are produced by varying the size of the silicon slots. Each neuron value is approximated by three identical slots, and the distance between two adjacent hidden layers is 300 μm. The photonic integrated DONN architecture can perform complicated functions at the speed of light with low power consumption due to its natural intrinsic character. Additionally, the manufacturing process of the chip is compatible with the CMOS process, which is convenient for large-scale and low-cost manufacturing. Furthermore, compared with other ONNs, our presented DONN has the advantages of simple structure design, all-optical passive operation, and massive scale neuron integration, etc. This deep learning framework may facilitate other applications includes speech recognition, data mining, object classification, and so on.

**Funding.** National Key Research and Development Program of China (2019YFB1803500); National Natural Science Foundation of China (61771284, 62135009).

**Disclosures.** The authors declare no conflicts of interest.

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

## References

1. A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, (IEEE, 2013), pp. 6645–6649.
2. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," https://arxiv.org/abs/1406.1078.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Commun. ACM **25**(6), 1097–1105 (2012).
4. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis **42**(9), 60–88 (2017).
5. M. H. Shoreh, U. S. Kamilov, I. N. Papadopoulos, A. Goy, C. Vonesch, M. Unser, and D. Psaltis, "A learning approach to optical tomography," https://arxiv.org/abs/1502.01914.
6. Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," Optica **4**(11), 1437–1443 (2017).
7. K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," IEEE Trans. on Image Process. **26**(9), 4509–4522 (2017).
8. Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," Light. Sci. & Appl. **7**(2), 17141 (2018).
9. A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," Optica **4**(9), 1117–1125 (2017).
10. K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," Magn. Reson. Med **79**(6), 3055–3071 (2018).
11. Y. Rivenson, H. Ceylan Koydemir, H. Wang, Z. Wei, Z. Ren, H. Gunaydın, Y. Zhang, Z. Gorocs, K. Liang, D. Tseng, and A. Ozcan, "Deep learning enhanced mobile-phone microscopy," ACS Photonics **5**(6), 2354–2364 (2018).

Optics EXPRESS


12. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
13. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**(6406), 1004–1008 (2018).
14. T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, "Fourier-space diffractive deep neural network," Phys. Rev. Lett. **123**(2), 023901 (2019).
15. T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," Optica **5**(7), 864–871 (2018).
16. D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–14 (2020).
17. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**(7755), 208–214 (2019).
18. E. Khoram, A. Chen, D. Liu, L. Ying, Q. Wang, M. Yuan, and Z. Yu, "Nanophotonic media for artificial neural inference," Photonics Res. **7**(8), 823–827 (2019).
19. J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, "Reinforcement learning in a large-scale photonic recurrent neural network," Optica **5**(6), 756–760 (2018).
20. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," Optica **6**(9), 1132–1137 (2019).
21. M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, "All-optical nonlinear activation function for photonic neural networks," Opt. Mater. Express **8**(12), 3851–3863 (2018).
22. I. A. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable electro-optic nonlinear activation functions for optical neural networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–12 (2020).
23. M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," Opt. Express **27**(10), 14009–14029 (2019).
24. S. Zarei, M.-r. Marzban, and A. Khavasi, "Integrated photonic neural network based on silicon metalines," Opt. Express **28**(24), 36668–36684 (2020).
25. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Sci Rep **8**, 1–10 (2018).
26. M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger, "Massively parallel amplitude-only fourier neural network," Optica **7**(12), 1812–1819 (2020).
27. C. Qian, X. Lin, X. Lin, J. Xu, Y. Sun, E. Li, B. Zhang, and H. Chen, "Performing optical logic operations by a diffractive neural network," Light Sci Appl **9**(1), 59 (2020).
28. Z. Wu, M. Zhou, E. Khoram, B. Liu, and Z. Yu, "Neuromorphic metasurface," Photonics Res. **8**(1), 46–50 (2020).
29. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol. **64**(5), 304–310 (1989).
30. S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (elm) technique for heart disease diagnosis," in *2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015)*, (IEEE, 2015), pp. 1–3.
31. K. Saxena Purushottam and R. Sharma, "Efficient heart disease prediction system," Procedia Computer Science **85**, 962–969 (2016).
32. S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis," Physica A: Statistical Mechanics and its Applications **482**(15), 796–807 (2017).
33. J. Vayashree and H. P. Sultana, "A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier," Program Comput Soft **44**(6), 388–397 (2018).
34. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mob. Inf. Syst. **2018**(8), 1–21 (2018).
35. S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, "Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowl Inf Syst **58**(1), 139–167 (2019).
36. Z. Wang, T. Li, A. Soman, D. Mao, T. Kananen, and T. Gu, "On-chip wavefront shaping with dielectric metasurface," Nat. Commun. **10**(1), 1–7 (2019).
37. D. H. Raguin and G. M. Morris, "Antireflection structured surfaces for the infrared spectral region," Appl. Opt. **32**(7), 1154–1167 (1993).
38. S. Rytov, "Electromagnetic properties of a finely stratified medium," Sov. Phys. JEPT **2**, 466–475 (1956).