Optical and Quantum Electronics (2026) 58:239
https://doi.org/10.1007/s11082-026-08835-y

A novel deep learning approach for accurate and efficient design of LNOI power splitters

Huriye Gencal[1,2] · Abdullah Aksoy[2] · Enes Yigit[2,3] · Umut Aydemir[2,3] · Mustafa Demirtas[2,3]

Received: 9 January 2026 / Accepted: 14 April 2026

Abstract
Photonic Integrated Circuits (PICs), owing to their high speed, low power consumption, and compact structure, lie at the core of modern optoelectronic technologies. The design of these circuits requires high accuracy and intensive computational cost. In this study, a novel Deep Neural Network (DNN)-based framework is proposed for designing and predicting the performance of arbitrary-ratio power splitters on the Lithium Niobate on Insulator (LNOI) platform. A dataset constructed using fundamental geometric parameters such as width, height, length, and auxiliary dimensions was processed with the proposed DNN model, yielding high prediction accuracy. The model achieved strong agreement in the training, validation, and testing stages, with R^2 values of 0.95, 0.97, and 0.97, respectively. The corresponding error metrics were RMSE = 3.08, 2.4, and 2.5, and MAPE = 4.02%, 3%, and 3.1%, respectively. Extensive analyses across various epoch numbers (500–10,000), batch sizes (2–64), and optimizers (Adam, SGD, RMSProp) revealed that the Adam optimizer, with 5,000 epochs and a batch size of 64, achieved the optimal balance between accuracy, convergence speed, and generalization. Furthermore, a detailed analysis of the influence of input parameters on outputs revealed that L_1 and W were the most critical factors. The trained model was also validated on an independent dataset from the literature, demonstrating excellent generalization ability with $R = 0.991$, RMSE = 1.98, and MAPE = 3.42%. To facilitate practical use of the proposed framework, an interactive MATLAB application was developed, enabling both forward prediction of power-splitting ratios from user-defined geometric inputs and inverse design of optimal parameters corresponding to a target output ratio through an integrated DNN–optimization workflow. This tool significantly accelerates device evaluation and design-space exploration, making the methodology readily applicable to real-world photonic design tasks. These results indicate that the proposed approach not only accelerates the design process but also enhances the understanding of input-output relationships, thereby providing a reliable methodology for photonic device optimization.

Keywords Lithium niobate on insulator (LNOI) · Power splitter · Deep neural networks (DNN) · Integrated photonics

Extended author information available on the last page of the article

Published online: 22 April 2026

1 Introduction

Photonic integrated circuits (PICs) have emerged as a key technology owing to their advantages of high bandwidth, low power consumption, and compact integration, enabling applications ranging from communications to quantum information processing. The much faster propagation of light compared to electrons allows these systems to outperform electronic counterparts, particularly in data centers and long-haul communication networks (Jalali and Fathpour 2006; Reed et al. 2010). The integration of waveguides and optical components on a single chip through nanofabrication techniques enables the realization of large-scale photonic networks and underscores the scalability of the platform (Bogaerts et al. 2012). Moreover, the ability to simultaneously transmit different wavelengths and polarizations enhances the capacity through wavelength and polarization-division multiplexing (WDM and PDM) techniques, establishing PICs as a critical component of modern communication infrastructures (Anderson and Webster 2016; Novick et al. 2023; Xu et al. 2005). With these advantages, PICs stand out as a powerful platform for next-generation solutions in communication, sensing, biophotonics, and quantum technologies. Among the various material platforms, silicon-on-insulator (SOI) has long played a dominant role in this field. SOI has enabled the design of compact and cost-effective photonic circuits owing to its high refractive index contrast and compatibility with CMOS fabrication processes (Soref 2010; Vivien and Pavesi 2016). With its high refractive-index contrast and CMOS-process compatibility, the SOI platform has enabled compact, cost-efficient photonic circuits (Soref 2010; Vivien and Pavesi 2016). Building on these advantages, devices on this platform are widely adopted in wavelength-division multiplexing (WDM) communication systems (Jalali and Fathpour 2006). Nevertheless, silicon photonics technology still faces certain limitations. The absence of second-order (χ^2) nonlinearity due to the centrosymmetric crystal structure of silicon (Soref 2006), together with the restricted electro-optic modulation capacity and the conflicting requirements among modulation speed, bandwidth, low loss, power consumption, and CMOS compatibility (Reed et al. 2010), present significant challenges in this field. Consequently, alternative photonic platforms, particularly lithium niobate and other emerging dielectric materials, have attracted growing attention in recent research (Poberaj et al. 2012; Wang et al. 2018). In this context, one of the most promising candidates is the thin-film lithium-niobate-on-insulator (LNOI) platform. LNOI consists of a sub-micron-thick lithium niobate (LN) layer, a buried silica layer, and either an LN or silicon substrate. This structure enables strong optical confinement while preserving the intrinsic advantages of LN. Owing to its high electro-optic coefficients ($r_{33} \approx 30$ pm/V), pronounced second-order nonlinear response, low propagation losses, and broad transparency window (0.35–5 μm), LNOI has demonstrated outstanding performance in a wide range of devices, including modulators, frequency converters, integrated lasers, quantum light sources, and detectors (Boes et al. 2023; Feng et al. 2024; Hang and Ang 2017; Poberaj et al. 2012).

In photonic integrated circuits, power splitters serve as fundamental building blocks, distributing optical power from a single input waveguide into multiple outputs with predetermined ratios, thereby enabling key functionalities such as interferometers, sensing systems, communication circuits, and quantum optics applications. Conventional power splitters rely on three main principles: multimode interference (MMI) structures (Soldano and Pennings 1995), Y-branch configurations (Chung et al. 2006; Rangaraj et al. 1989), and directional couplers (Takahashi and Nonaka 1977; Yariv 1973). On the SOI platform,

devices based on these structures have been demonstrated with high efficiency (Bogaerts et al. 2012). In contrast, LNOI-based technologies provide opportunities for designing more compact, low-loss, and reconfigurable power splitters. In recent years, particular attention has been devoted to the class of arbitrary-ratio power splitters. Moving beyond the commonly employed 50:50 ratio, these devices enable power distribution at ratios of 60:40, 70:30, or other proportions, thereby enhancing flexibility in photonic circuit design. Such devices play an important role in applications including optical neural networks (Shen et al. 2017), programmable interferometric circuits (Bogaerts et al. 2020), communication systems (Jalali and Fathpour 2006), and polarization management (Dai et al. 2013). To realize these functionalities, various approaches have been reported in the literature. These include asymmetric MMI structures (Deng et al. 2014), shape and topology optimization techniques (Frellsen et al. 2016; Liao et al. 2024), adiabatic rib waveguides (Liu et al. 2024), inverse-designed Si_3N_4 splitters (Song et al. 2024), and multi-output arbitrary-ratio configurations (Liu et al. 2025). More recently, such designs have also been successfully adapted to the LNOI platform (Lin et al. 2023; Lyu et al. 2025; Yousefi et al. 2025).

In recent years, machine learning (ML) methods have been increasingly integrated into design processes. Large-scale datasets obtained from electromagnetic solvers are utilized to develop predictive models that establish direct correlations between device geometries and output performance (Malkiel et al. 2018; Ma et al. 2021; Peurifoy et al. 2018a). Compared to conventional trial-and-error-based optimization approaches, these methods reduce design time, provide quantitative evaluation of the impact of parameters on device performance, and enable systematic exploration of large design spaces (Molesky et al. 2018; Pan and Pan 2023).

In particular, data-driven approaches, such as Neural Networks (NNs), allow for behavioral prediction even in cases where theoretical models are unavailable (Formisano and Tucci 2024). NN-based methods have previously been used to study specific photonic devices, such as phase-shifting structures (Liu et al. 2018), photonic crystal nanocavities (Asano and Noda 2018), and nanophotonic particles (Peurifoy et al. 2018b). NNs offer powerful applications in this field by accelerating design processes and solving complex optimization problems. For instance, Yiğit et al. developed an artificial neural network (ANN)– based model that enables MEMS diaphragm analysis without resorting to complex and time-consuming finite element method (FEM) procedures (Yigit et al. 2022). In another study, an artificial neural network developed for simulating the behavior of nanophotonic particles and inverse design successfully replicated the light scattering behavior of multilayer particles, thereby achieving simulation times several times faster than traditional methods. The trained network efficiently solved inverse design problems by using backpropagation for analytical gradient computation (Peurifoy et al. 2018b).

Despite these advances, the use of NNs in photonic device design remains limited in capturing the highly complex and nonlinear relationships between device geometries and optical responses. To overcome this limitation, Deep Neural Networks (DNNs) with layered architectures are employed. Takashi Asano and Susumu Noda demonstrated that deep learning approaches can be effectively utilized to optimize the Q factor of photonic crystal nanocavities, achieving high values through data-driven optimization of cavity geometries. A trained neural network, capable of rapidly estimating the gradient of the Q factor, enabled optimization over wide parameter spaces that are typically inaccessible to conventional methods, achieving an exceptionally high Q factor of 1.58×10^9 (Asano and Noda 2018).

In another example, a design methodology combining a deep learning model with gradient descent achieved much faster convergence than traditional optimization techniques in the direct design of arbitrary-ratio photonic power splitters (Kong et al. 2024; Yigit et al. 2022).

Previous studies on the design of arbitrary-ratio power splitters on the LNOI platform have predominantly relied on conventional methods or inverse optimization techniques (Shen et al. 2024; Zhu et al. 2021). However, these approaches often involve high computational costs and are unable to fully capture the complex and nonlinear relationships between geometric parameters and optical responses (Kang et al. 2024). To address these limitations, this study introduces a deep learning–based framework that effectively captures such dependencies, providing a robust, data-driven methodology that accelerates photonic device simulations, facilitates inverse design, and substantially reduces computational burden compared to traditional optimization techniques.

In this study, a Deep Neural Network (DNN)-based framework is introduced for the prediction and inverse design of arbitrary-ratio power splitters on the Lithium Niobate on Insulator (LNOI) platform. Unlike existing deep-learning-based photonic inverse design approaches that predominantly rely on black-box optimization or topology-driven, pixel-level representations, the proposed framework establishes a geometry-aware, data-driven surrogate model that directly maps six physically interpretable geometric parameters (Width, h_1, h_2, L, W, and L_1) to the output power-splitting ratios. A key distinction of this work lies in the unified integration of high-accuracy forward prediction and gradient-based inverse design within a single computational workflow. The trained DNN, comprising approximately 16,000 learnable parameters, serves not only as a compact and efficient forward surrogate model (achieving $R^2 \approx 0.97$ and RMSE$\approx$2.4), but is also explicitly embedded into a gradient-based optimization loop to solve the inverse design problem without requiring repeated electromagnetic simulations, thereby significantly reducing computational overhead. In contrast to many prior studies where inverse design is treated as a separate or post-processing task, the proposed approach enables continuous and real-time interaction between prediction and optimization stages.

Beyond performance prediction, the framework incorporates a quantitative sensitivity and parameter importance analysis, revealing that the L_1 parameter alone accounts for approximately 66% of the output variance, followed by the width W as the second most influential variable; this level of physical interpretability is largely absent in existing DNN-based photonic design studies despite their reported accuracy. Another distinguishing feature of this work is the systematic external validation of model generalization using independent datasets extracted from the literature, yielding $R = 0.991$, RMSE = 1.98, and MAPE = 3.42%, which quantitatively demonstrates robust transferability across different LNOI splitter geometries and design regimes an evaluation step rarely reported in prior deep-learning-driven photonic inverse design studies. To further bridge the gap between theory and practical deployment, an interactive MATLAB-based application has been developed that seamlessly integrates forward prediction and inverse design functionalities under fabrication-constrained parameter ranges. This user-oriented tool enables rapid evaluation of arbitrary geometries and efficient computation of optimal design parameters for target power-splitting ratios, positioning the proposed framework not merely as a high-accuracy regression model, but as a practical, interpretable, and generalizable design methodology for accelerating the optimization of LNOI-based arbitrary-ratio power splitters.

2 Material & method

This study comprises two main stages, namely the construction of the dataset and the implementation of the proposed DNN model. A detailed block diagram illustrating the processes in each stage is presented in Fig. 1.

The block diagram in Fig. 1 provides a two-stage summary of the study. In the dataset construction stage, the design variables Width, h_1, h_2, L, W, and L_1 were employed as input parameters. The optimal value ranges of these parameters were determined, and subsequently, the S-parameter results were divided into representative percentage ratios (50–50, 60–40, 70–30, 80–20) and visualized. Simulations were then performed for each ratio, generating a large volume of data. In the proposed artificial intelligence stage, training, validation, and testing were initially carried out using standard machine learning algorithms (Support Vector Machine (SVM), Gaussian Process Regression (GPR), Tree, Neural Network (NN), etc.), followed by the application of the newly proposed DNN model to the same dataset. After training, the model was stored and subsequently used both to validate its performance against literature-based datasets and to reveal the dependencies between input and output parameters. Through the recorded DNN network, the effect of each input parameter on the outputs was analyzed individually, allowing the identification of the variables

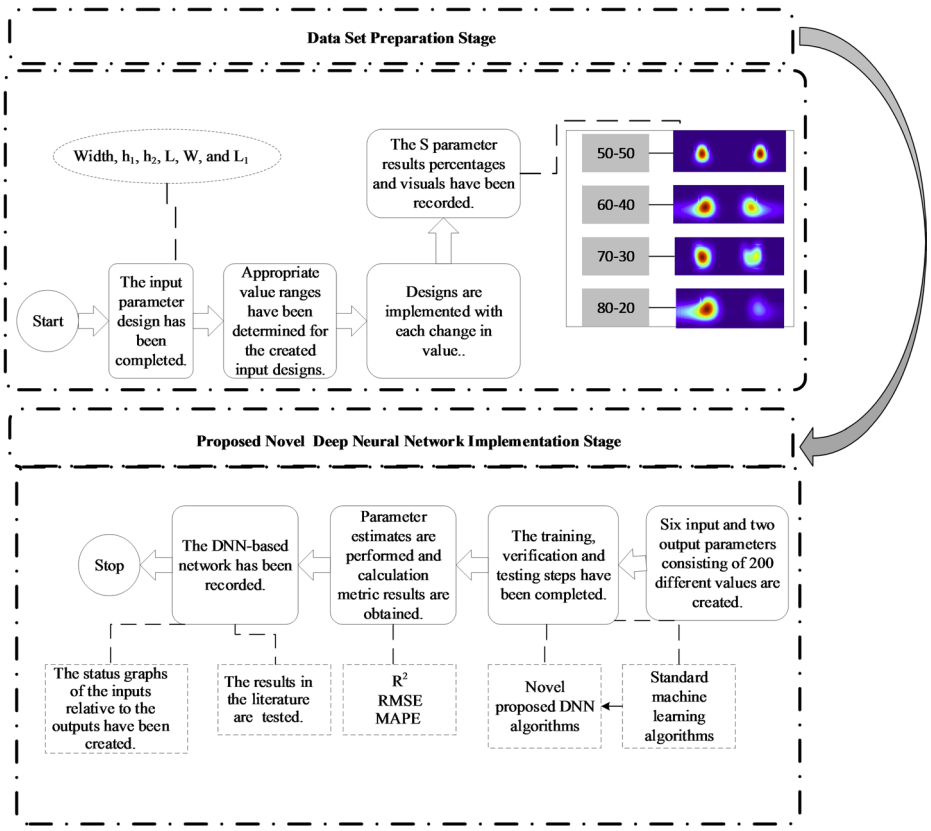


Fig. 1 The block diagram of the conducted study

that exert the strongest influence. In this way, both performance comparisons and a detailed examination of input–output parameter relationships were accomplished.

2.1 Design and simulations

The design and analysis of arbitrary-ratio power splitters were carried out on the lithium niobate on insulator (LNOI) platform. The refractive indices of the materials were defined based on literature data, with $LiNbO_3$ as the guiding layer and SiO_2 as the buried oxide layer. The target operating wavelength was fixed at 1550 nm, a standard wavelength for optical communication systems.

In our study, the Eigenmode Expansion (EME) method was employed for analyzing optical waveguide-based devices with long propagation distances, such as the designed power splitter. This method employs a fully vectorial and bi-directional numerical approach that solves Maxwell's equations in the frequency domain, allowing accurate mode characterization of photonic structures. This approach offers a more efficient and faster solution compared to alternative techniques, such as the Finite-Difference Time-Domain (FDTD) method, which can be computationally expensive, particularly for long and tapered structures. The fundamental principle of the EME method is to decompose optical fields into a superposition of locally computed eigenmodes. This decomposition is essential as it facilitates the calculation of the Scattering Matrix (S-Matrix), which regulates the optical transitions across the entire device. In numerical analyses, Perfectly Matched Layer (PML) boundary conditions were applied to suppress non-physical reflections that may occur at the simulation boundaries and to ensure accurate mode solving.

Figure 2 (a–c) illustrates the total geometry and layer configuration of the proposed rib-type waveguide structure. Figure 2a presents an isometric view of the multilayer design implemented on a lithium niobate-on-insulator (LNOI) platform. As seen in the figure, the uppermost layer is the waveguide core, followed by the SiO_2 insulating layer, and finally the high-resistivity silicon substrate. The waveguide region is formed as a raised rib structure through an etching process applied to the lithium niobate film, which enables effective optical confinement both laterally and vertically. Figure 2b shows the cross-sectional view of the waveguide, along with its dimensional parameters. In this configuration, h_1 denotes the slab height, while h_2 represents the rib height of the waveguide. The rib height enhances vertical mode confinement, whereas the slab thickness limits downward leakage and ensures single-mode operation. The width parameter corresponds to the rib top width, which must be precisely optimized to maintain single-mode propagation (Gencal et al. 2025, Demirtas et al. 2016). Figure 2c depicts the top view of the design, including the primary dimensional parameters from input to output. Here, L represents the total optical propagation length, W the width of the multimode region, and L_1 the length of the rectangular section that is removed. Asymmetric output power distributions can be obtained by changing L_1. The width, h_1, h_2, L, L_1, and W parameters defined in Fig. 2. (a–c) have an important role in the fine tuning of the power splitting ratios at the output ports, and this relationship has been verified by Lumerical MODE/EME-based simulations.

The output power distribution was calculated using the output powers derived from the squared magnitudes of the S-parameters ($|S|^2$) obtained in Lumerical EME. The upper and lower port powers are denoted as O_1 and O_2, respectively. The percentage contribution of each port to the total output power is expressed as:

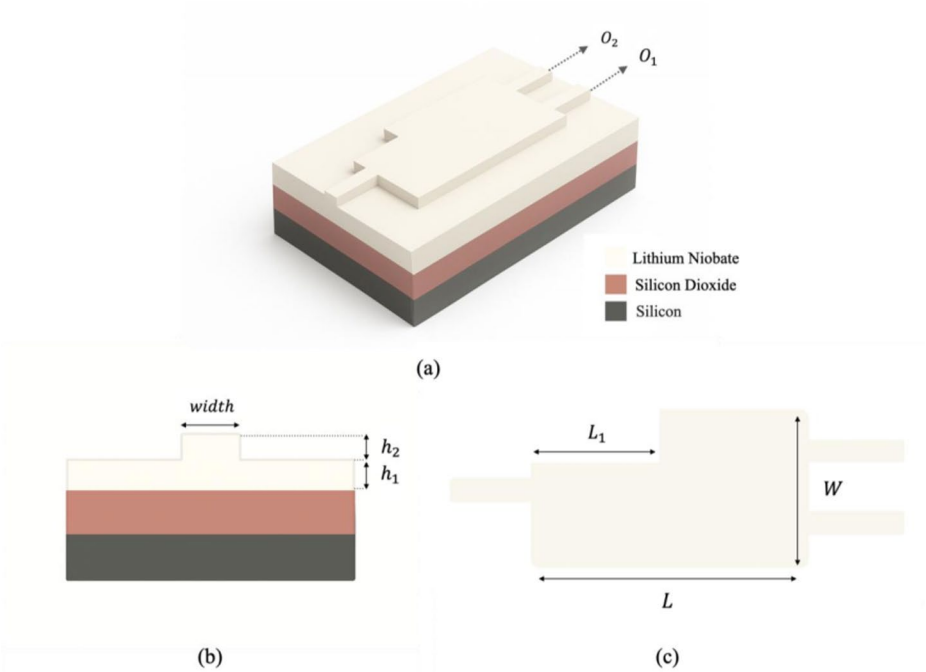


Fig. 2 LNOI-based power splitter design and geometric parameters: **a** 3D isometric view, **b** cross-sectional view, **c** top view

$$P_1 = \frac{O_1}{O_1 + O_2} \times 100 \tag{1}$$

$$P_2 = \frac{O_2}{O_1 + O_2} \times 100 \tag{2}$$

where P_1 denotes the percentage power contribution of the upper port (O_1), and P_2 represents that of the lower port (O_2). From this definition, it directly follows that $P_1 + P_2 = 100$. In this way, the percentage power distribution at the output ports has been quantitatively characterized as a function of the device geometry.

Figure 3(a–f) illustrates the percentage power distributions at the output obtained from different design parameters of the splitter and presents representative field patterns used during the construction of the simulation dataset, which consists of a total of 200 samples generated using Lumerical MODE/EME simulations. In Fig. 3(a), both output ports exhibit similar field intensities, corresponding to an approximately 50:50 power distribution. By systematically varying the geometric parameters within physically feasible and fabrication-constrained ranges, power splitting ratios of 60:40, 70:30, 80:20, 90:10 are obtained, as shown in Figs. 3(b–e), respectively. These representative configurations correspond to the five target power-splitting classes included in the dataset 50:50, 60:40, 70:30, and 80:20, 90:10 which were generated in a balanced manner to ensure uniform coverage of the design space. When the power ratio exceeds 60:40, the field intensity in the secondary port gradually decreases, indicating a reduced coupling contribution from that branch. This behavior

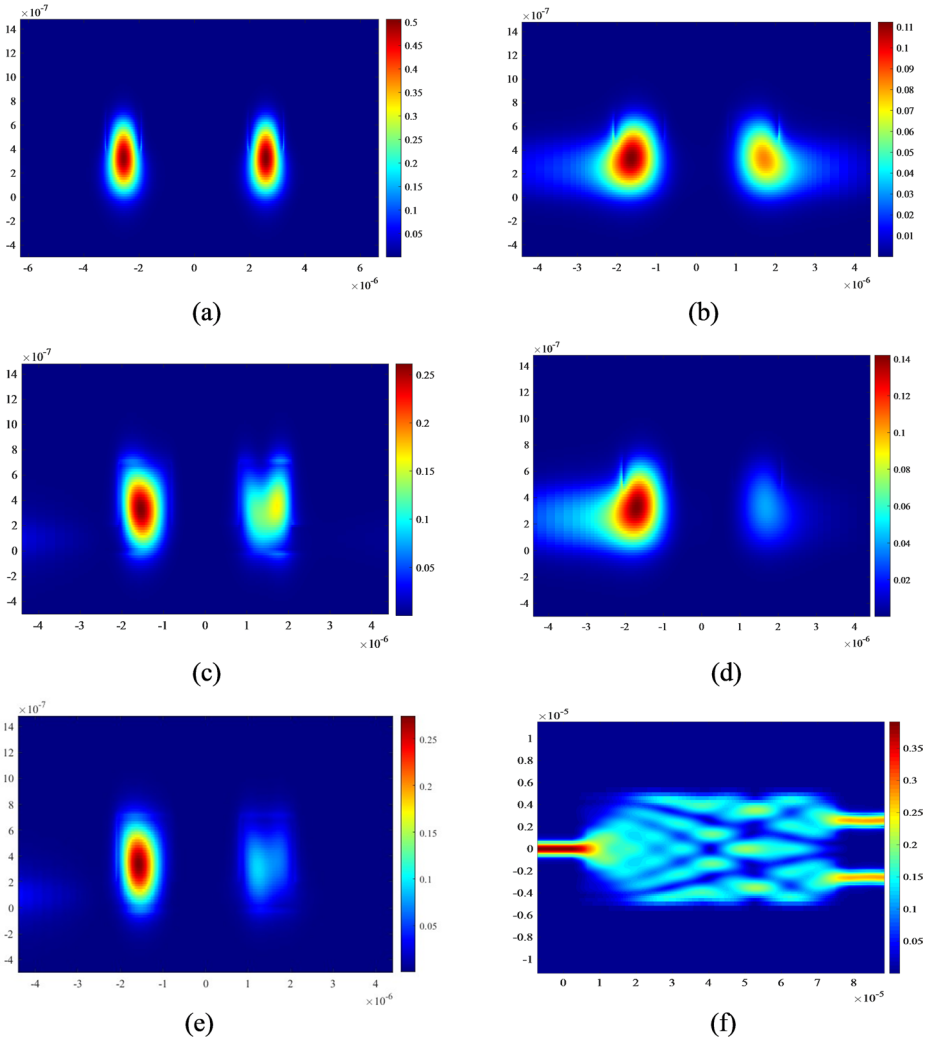


Fig. 3 Power distributions at the output of the splitter for different geometric parameters: **a** 50:50, **b** 60:40, **c** 70:30, **d** 80:20, **e**. 90:10, **f** Top-view propagation of 50:50 light distribution

demonstrates the ability of the proposed waveguide structure to precisely control the output power distribution through geometric tuning and forms the physical basis of the labeled samples used for training and validating the data-driven model. Figure 3(f) presents the top view of the optical structure corresponding to a 50:50 power splitter, illustrating a fundamental configuration in photonic integrated circuits in which the optical signal enters from a single input waveguide, propagates through the splitter region, and is evenly divided into two output waveguides; this configuration represents a typical reference case employed during dataset generation and physical verification of the model. To verify the reproducibility of the simulation dataset, three consecutive EME simulations were performed under identical geometric parameters for a representative test sample. All runs yielded identical

S-parameter matrices and power splitting ratios, confirming the deterministic nature of the Lumerical EME solver and the reliability of the dataset used for DNN training.

2.2 Proposed DNN model

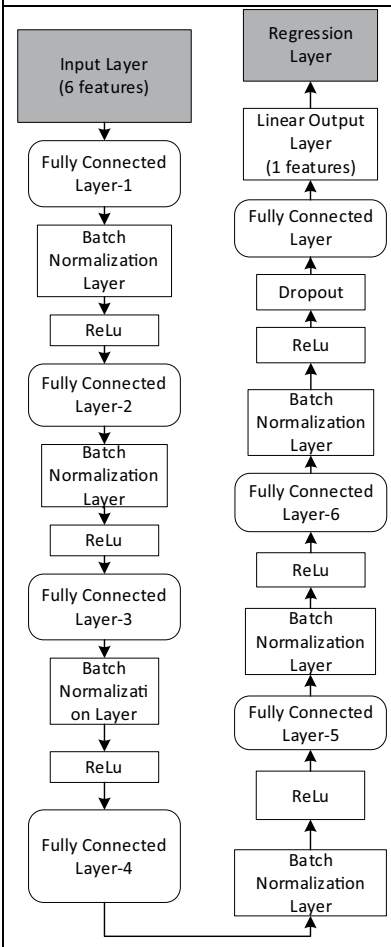
In the first stage of the study, standard machine learning algorithms (Support Vector Machine (SVM), Gaussian Process Regression (GPR), Tree, Neural Network (NN)) were employed to carry out the training, validation, and testing processes. However, these methods failed to adequately capture the complex and nonlinear relationships between input and output parameters. In particular, these approaches proved insufficient in modeling the influence of design parameters on the outputs, thereby failing to achieve the desired performance. Consequently, in the second stage, a specifically designed multilayer DNN architecture was developed.

The layers developed for the DNN architecture include Fully Connected Layers, ReLU, Batch Normalization, and Dropout. The fully connected layer performs a linear transformation of the data by directly associating all features in the input vector with each neuron, mathematically expressed as a = Weights × x + Bias. Here, the weight matrix and bias vector constitute the fundamental parameters updated during the learning process, thereby defining the input–output relationships of the layers. Following this transformation, the batch normalization layer normalizes the data by using the mean and variance values computed for each mini-batch during training. This reduces distributional shifts, enables a more stable learning process, and allows for the use of higher learning rates. The subsequent ReLU activation function introduces nonlinearity, enhancing the model's ability to learn more complex relationships. Defined as h = max(0, y), this function sets negative inputs to zero while transmitting positive values directly to the output. Finally, the dropout regularization technique is employed, where specific neurons are randomly deactivated during training to prevent the model from becoming overly dependent on particular features. This reduces the risk of overfitting and contributes to a more generalizable structure. The comprehensive architecture built using these layers is encapsulated in Table 1.

The proposed architecture consists of seven fully connected layers with a total of 16,065 learnable parameters. The first layer (Fully Connected Layer 1) includes a weight matrix of size 6 × 128 and 128 biases, amounting to 896 parameters. This layer projects the input features into a higher-dimensional space, facilitating the extraction of fundamental attributes. The second layer (Layer 2) contains 128 × 64 weights and 64 biases, yielding 8,256 parameters, and compresses the representations from the previous layer to capture deeper relationships. The third layer (Layer 3), with 64 × 64 weights and 64 biases (4,160 parameters), supports the learning of mid-level features. As the network depth increases, the number of parameters is deliberately reduced. Layer 4 (64 × 32 weights + 32 biases, 2,080 parameters), Layer 5 (32 × 16 weights + 16 biases, 528 parameters), and Layer 6 (16 × 8 weights + 8 biases, 136 parameters) progressively condense the information. At the final stage, Layer 7 comprises only 8 weights and one bias, resulting in an output with a total of 9 parameters. Through this structure, the network processes complex, high-dimensional inputs and predicts the final output in its most compact form.

Each weight defines how the input features are transformed across the network, while the bias terms adjust the activation thresholds of neurons, thereby enabling the formation of nonlinear decision boundaries. For example, the 128 biases in the first layer allow each

Table 1 Novel designed DNN model: layer-wise learnable parameter counts

Novel Designed DNN model	Parameter Numbers		
	Layer Name	Learnable Parameters	Parameter Counts
	Fully Connected Layer-1	Weights:6x128 Bias:128	896
	Fully Connected Layer-2	Weights:128x64 Bias:64	8256
	Fully Connected Layer-3	Weights:64x64 Bias:64	4160
	Fully Connected Layer-4	Weights:64x32 Bias:32	2080
	Fully Connected Layer-5	Weights:32x16 Bias:16	528
	Fully Connected Layer-6	Weights:16x8 Bias:8	136
	Fully Connected Layer-7	Weights:8x1 Bias:1	9
	Total Parameter Count		16065

feature map to have a distinct offset, enhancing the model's flexibility during learning. Similarly, the bias terms in the intermediate and final layers contribute to representing more complex functions beyond weight-only dependencies.

ReLU (Rectified Linear Unit) activation was applied in all layers to prevent gradient vanishing and ensure efficient learning. Batch Normalization layers were employed to stabilize gradient flow, while Dropout regularization was used to reduce overfitting. The sequential integration of these layers enables the proposed model not only to achieve strong representational capacity but also to maintain stable and effective training performance. Mean Squared Error (MSE) is selected as the loss function, and one of momentum-based SGD, Adam, or RMSProp is used as the optimisation method.

3 Results and discussion

3.1 DNN-based prediction results

At this stage of the study, as illustrated in the block diagram in Fig. 1, the dataset constructed in the first phase was used for performance evaluation. To establish baseline results, several commonly used machine learning regression methods were implemented and tested. These methods were grouped into four categories: Support Vector Machines (SVM), Gaussian Process Regression (GPR), Tree-based regression models (Tree), and classical Neural Network (NN) models. Within each category, multiple model configurations were examined using different kernel functions, regression structures, and network sizes, resulting in a total of 25 distinct models. Specifically, the SVM group included linear, quadratic, cubic, and Gaussian kernel functions; the GPR group comprised exponential, squared exponential, rational quadratic, and Matérn covariance functions; Tree-based models were evaluated under coarse, medium, and fine regression settings; and NN models were implemented with different hidden neuron configurations (shallow and medium architectures).

In addition to the traditional machine learning baselines, three lightweight deep learning architectures were implemented and benchmarked to provide a more comprehensive evaluation of the proposed DNN architecture: a simplified 1D Convolutional Neural Network (CNN) comprising two convolutional blocks followed by fully connected layers (~52,000 parameters), a compact Transformer-based regression model with a single encoder block incorporating multi-head self-attention (~21,000 parameters), and a lightweight ResNet with residual skip connections adapted for tabular regression (~31,000 parameters).

All models were trained and evaluated using the same dataset and an identical data partitioning strategy, with 70% of the data used for training, 15% for validation, and 15% for testing. Input features were processed using the same normalization procedure prior to training. The Adam optimizer was applied with identical hyperparameter settings and early stopping criteria across all models to ensure a fair comparison. For each traditional model category, the configuration yielding the highest validation R^2 and the lowest RMSE was selected as the representative baseline for comparison with the proposed DNN model.

Figure 4 illustrates the regression performance of the proposed DNN-based model obtained for different epoch numbers ranging from 500 to 10,000. During training, five different batch sizes (2, 8, 16, 32, and 64) were examined; however, only the results corresponding to a batch size of 64 are presented, as this configuration exhibited the most stable convergence behavior across the investigated epoch range. In addition, three optimization algorithms Adam, SGD, and RMSProp were evaluated, among which the Adam optimizer provided faster convergence and lower validation loss, resulting in a more stable training process.

As observed in Fig. 4, increasing the number of epochs leads to a pronounced reduction in both training and validation losses, as well as RMSE values. At 500 epochs, the model yields relatively high error levels (validation RMSE = 31.33, R^2 = 0.72), while a substantial improvement is observed around 2000 epochs (RMSE = 6.87, R^2 = 0.95). The minimum validation error is achieved at approximately 5000 epochs (RMSE = 3.73, R^2 = 0.97), indicating a balanced operating region between learning capacity and generalization performance. For higher epoch values, particularly around 10,000 epochs, a slight degradation in validation performance is observed (R^2 = 0.91), suggesting the onset of overfitting.

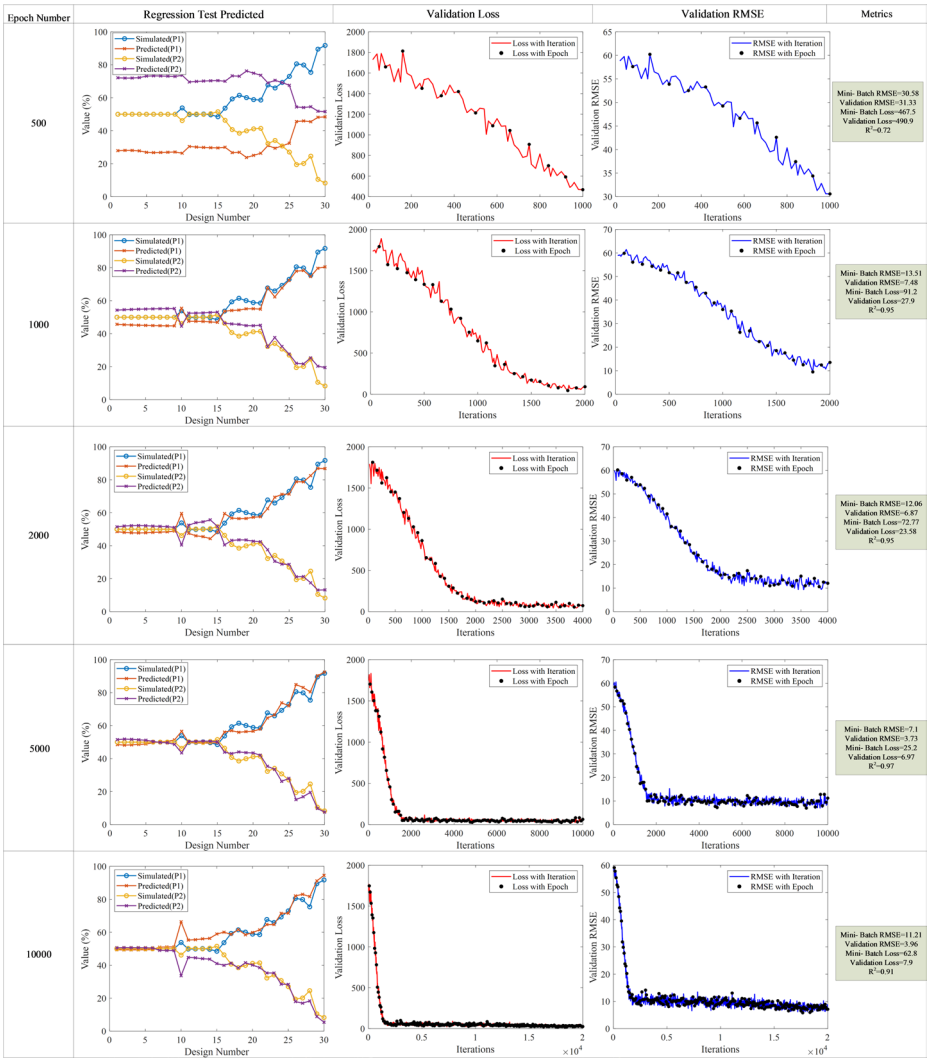


Fig. 4 Regression results for DNN-based design

To address this behavior, a validation-based early stopping strategy was employed during training, allowing the optimization process to terminate once no further improvement in validation performance was detected.

In addition, to assess the robustness of the network design and mitigate overfitting, a limited ablation study was conducted focusing on three key aspects: network depth (number of fully connected layers), model capacity (total number of learnable parameters), and the presence of dropout regularization. Architectures with fewer layers or significantly reduced parameter counts were found to be insufficient for capturing the highly nonlinear relationship between geometric parameters and output power ratios, while larger-capacity models did not yield consistent improvements in validation performance. Based on these observa-

tions, a network configuration comprising approximately 16,000 learnable parameters was selected as a balanced trade-off between representational capacity and generalization. A dropout rate of 0.1 was selected for the final model, as it preserved training stability while effectively constraining validation error without impairing convergence.

Furthermore, dropout regularization was incorporated into the network architecture, and different dropout settings were examined. A dropout rate of 0.1 was selected for the final model, as it preserved training stability while effectively constraining validation error without impairing convergence.

In light of these findings, the configuration employing the Adam optimizer, a batch size of 64, a dropout rate of 0.1, and a training process terminating at approximately 5000 epochs was identified as the most balanced setup in terms of prediction accuracy and generalization capability. The best results obtained for each method are reported in Table 2.

Following this analysis, Table 2 presents the quantitative comparison of all tested methods. Among the traditional machine learning approaches, the NN model achieved the highest test R^2 of 0.85, while Tree, SVM, and GPR yielded test R^2 values of 0.84, 0.82, and 0.81, respectively, with RMSE values ranging between 6.9 and 7.8. The three lightweight deep learning architectures demonstrated moderate improvements over the traditional baselines in training and validation phases; however, their test performance remained within a comparable range: ResNet achieved a test R^2 of 0.84 (RMSE = 6.88, MAPE = 8.37%), CNN reached 0.83 (RMSE = 6.94, MAPE = 8.41%), and Transformer yielded 0.82 (RMSE = 7.14, MAPE = 8.10%). Notably, despite containing more parameters than the proposed DNN — approximately 52,000 for CNN, 31,000 for ResNet, and 21,000 for Transformer — none of these architectures provided a meaningful accuracy advantage over the traditional baselines for this low-dimensional tabular regression problem, suggesting that architectural complexity alone is insufficient to capture the nonlinear input–output dependencies present in the dataset.

In contrast, the proposed DNN achieved a training R^2 of 0.94, a validation R^2 of 0.95, and a test R^2 of 0.90, with a validation RMSE of 3.32 and a training MAPE of 4.01%. This represents a reduction of approximately 50–55% in RMSE and a roughly twofold improvement in MAPE during training compared to all baseline methods, confirming the superior capability of the proposed architecture in modeling complex nonlinear relationships with high predictive reliability. Furthermore, the proposed DNN achieves this performance with only 16,065 learnable parameters — fewer than any of the lightweight deep learning architectures evaluated demonstrating that the proposed design is not only the most accurate but also the most parameter-efficient solution among all compared methods.

As seen in Table 3, the traditional machine learning methods (Tree, SVM, GPR, and NN) achieve extremely fast inference times in the range of 0.1–0.8 ms owing to their simple computational structures; however, as demonstrated in Table 2, this speed advantage comes at the cost of significantly lower prediction accuracy. Among the lightweight deep learning architectures, the Transformer model exhibits the slowest inference time (~8–12 ms) despite its relatively moderate parameter count (~21,000). This is attributable to the computational overhead of the self-attention mechanism, which is inherently less suited for low-dimensional tabular regression tasks involving only six scalar input features. The 1D-CNN (~52,000 parameters, 3–5 ms) and ResNet (~31,000 parameters, 4–6 ms) require considerably more parameters than the proposed DNN while delivering lower prediction accuracy across all evaluation phases, as confirmed by Table 2. The proposed DNN, with

Table 2 Evaluation Metrics

Methods	Training Results			Validation Results		
	R^2	MAPE (%)	RMSE	R^2	MAPE (%)	RMSE
Tree	0.71	8.8	6.70	0.72	8.72	7.5
SVM	0.78	8.76	7.1	0.78	8.7	7.2
GPR	0.792	8.6	7.06	0.78	8.6	7.1
NN	0.81	6.8	5.95	0.8	6.8	5.7
CNN	0.79	7.90	6.41	0.80	7.85	6.30
Transformer	0.77	8.20	6.68	0.78	8.15	6.55
ResNet	0.80	7.65	6.18	0.81	7.70	6.05
Proposed DNN	**0.94**	**4.01**	**3.34**	**0.95**	**4.25**	**3.32**
Testing Results						
R^2		MAPE (%)				RMSE
0.84		8.36				6.9
0.82		8.48				7.12
0.81		9.2				7.8
0.85		8.3				6.8
0.83		8.41				6.94
0.82		8.10				7.14
0.84		8.37				6.88
0.90		**8.10**				**6.30**

only 16,065 learnable parameters, achieves the fastest inference time among all deep learning models (1–3 ms) while simultaneously delivering the highest prediction accuracy. This combination of the lowest parameter count, the fastest inference speed, and the highest accuracy among all deep learning architectures evaluated in this study collectively confirms the design rationality and practical efficiency of the proposed model for LNOI power splitter performance prediction.

Regarding computational efficiency relative to direct electromagnetic simulation, all data-driven methods provide a speed-up of at least four to five orders of magnitude over a single Lumerical EME simulation, which requires approximately 9–11 min per design point on the same workstation. The proposed DNN achieves a speed-up factor of approximately 180,000–660,000× for a single forward evaluation, while even the slowest deep learning model (Transformer, ~8–12 ms) provides a speed-up of approximately 45,000–82,500×. This acceleration is of particular practical significance in design-space exploration tasks, where hundreds or thousands of candidate geometries must be evaluated; a task that would require days or weeks of continuous simulation time using the EME solver can be completed within seconds using the proposed DNN-based framework, confirming the substantial engineering practicality of the developed tool.

During the testing phase of the study, 30 distinct test samples, corresponding to 15% of the dataset, are presented in Table 4. The table includes the input parameters (Width, h_1, h_2, L, W, L_1) and the output variables P_1 and P_2. The P_1 and P_2 ratios represent complementary power distribution percentages that sum to one hundred, illustrating a comparative relationship between the simulated and DNN-predicted results.

When Fig. 5a and b are considered together, it is evident that the proposed DNN model demonstrates consistent predictive performance in terms of both agreement with the real values and error distribution. In the comparison shown in Fig. 5a, the predicted values closely follow the real values, and the model is also able to capture sudden variations successfully. Figure 5b presents the distribution of prediction errors, indicating that the majority of errors are concentrated at low levels, with only a limited number of noticeable deviations. This suggests that the model has effectively learned and generalized the input–output relationships, while the deviations arise from local differences rather than systematic trends. The comprehensive study of both figures validates that the suggested model demonstrates strong performance regarding accuracy and stability, accurately capturing the fundamental input-output relationships.

3.2 Correlation of design parameters on input/output performance

In this section of the study, the effect of each input parameter on the output was examined in terms of both proportional effect and inter-parameter relationships, and the results are presented in Fig. 6. The information presented in Fig. 6 is based on the P1 output. Since the output parameters complement each other, the results obtained from one parameter are sufficient.

In Fig. 6a, the findings indicate that the input parameters exert varying degrees of influence on the model outputs. An analysis of the contribution percentages shows that h_1 accounts for 3.65%, h_2 for 3.65%, L for 7.76%, W for 18.70%, and L_1 for 66.20%, with L_1 demonstrating a dominant effect on the outputs. This distribution highlights the decisive role of the L_1 parameter in the model's learning of input–output relationships, while W

Table 3 Comparison of model complexity, inference speed, and computational efficiency relative to Lumerical EME simulation

Method	Parameter Count	Inference Time (per sample)	Speed-up vs. EME
Tree	—	~0.1 ms	~5,400,000×
SVM	—	~0.2 ms	~2,700,000×
GPR	—	~0.5 ms	~1,080,000×
NN	—	~0.8 ms	~675,000×
1D-CNN	~52,000	~3–5 ms	~108,000–220,000×
Transformer	~21,000	~8–12 ms	~45,000–82,500×
ResNet	~31,000	~4–6 ms	~90,000–165,000×
Proposed DNN	**16,065**	**~1–3 ms**	**~180,000–660,000×**
Lumerical EME	—	~9–11 min	Reference

emerges as a secondary influential variable, and the remaining parameters have relatively limited influence. Consequently, focusing on L_1 and W in design optimization studies is expected to provide higher accuracy and efficiency. In Fig. 6b, it is clearly observed that the relationship between the h_1 parameter and the model output is nonlinear. At the initial stage, the output value remains low, but as the parameter increases, the response rises considerably and reaches its maximum around the middle of the range. Beyond this optimal point, further increases in h_1 lead to a reduction in the output value. This behavior indicates that the system operates more effectively within a specific parameter range, whereas exceeding the threshold negatively affects performance. Therefore, this parameter does not influence the system response in a uniform manner but instead functions as a critical control variable by producing both positive and negative effects. In the case of the h_2 parameter shown in Fig. 6c, a nonlinear relationship similar to that of h_1 is observed. When the parameter takes small values, an increase in the output can be seen; however, beyond a certain threshold, this increase is replaced by a decline. This indicates that the effect of the parameter is bidirectional, improving the output within specific limits but negatively influencing system performance once the optimal value is exceeded. Hence, defining the optimal ranges of this parameter with precision is of critical importance for ensuring accurate model predictions and stable system operation. Figure 6d reveals a near-linear relationship between the parameter L and the output. As the parameter increases, the model output exhibits a consistent rise, indicating a sustained positive influence on system performance. The monotonic increase in the output confirms that the parameter is effectively captured by the model, with each incremental change at the input being directly reflected in the response. This behavior implies that enhancing the value of L during the design stage can systematically improve overall system performance. In Fig. 6e, an inverse relationship is observed between the parameter W and the output. At lower parameter values, the output remains at higher levels, whereas a progressive decrease in the output is observed as the parameter increases. This behavior indicates that increasing W reduces system performance and negatively influences the output. The consistently decreasing profile of the curve demonstrates that the model has identified this parameter as a strong negative determinant. Therefore, excessive enlargement of W weakens the system response, highlighting its role as a critical variable that must be constrained in optimization studies. In Fig. 6f, a much more pronounced and dominant relationship is observed for the L_1 parameter. The initially high output value decreases continuously and sharply as L_1 increases. This outcome indicates that L_1 has a predominant influence on the model output and that system performance is largely determined by this parameter.

Table 4 Test data results for DNN-based design

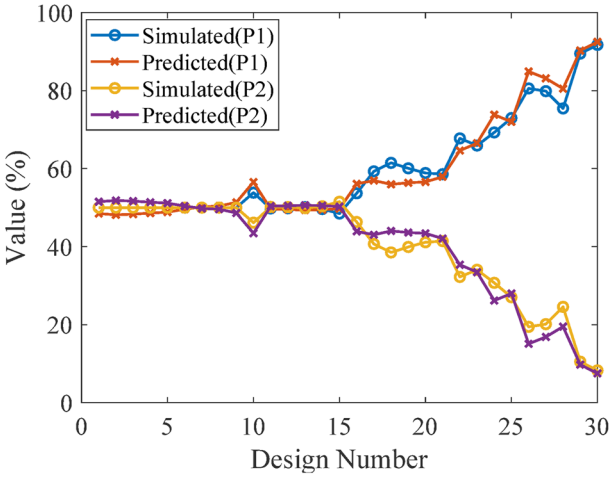
Design Number	Input Properties					
	Width (μm)	h_1 (μm)	h_2 (μm)	L (μm)	W (μm)	L_1 (μm)
1	1.25	0.2	0.5	26.4	7.6	0
2	1.25	0.2	0.5	26.4	7.9	0
3	1.25	0.2	0.5	26.4	8.2	0
4	1.25	0.2	0.5	26.4	8.5	0
5	1.25	0.2	0.5	26.4	8.8	0
6	1.25	0.2	0.5	26.4	9.1	0
7	1.25	0.2	0.5	26.4	9.4	0
8	1.25	0.2	0.5	26.4	9.7	0
9	1.25	0.2	0.5	26.4	10	0
10	1.25	0.2	0,5	26.4	9,5	3
11	1.25	0.5	0.2	75.54	10	3
12	1.25	0.5	0.2	75.54	10.5	3
13	1.25	0.5	0.2	75.54	11	3
14	1.25	0.5	0.2	75.54	11.5	3
15	1.25	0.5	0.2	75.54	12	3
16	1.25	0.2	0.5	75.54	13.5	0
17	1.25	0.4	0.3	26.4	9.5	3
18	1.25	0.4	0.3	26.4	10	3
19	1.25	0.5	0.2	26.4	8	3
20	1.25	0.5	0.2	26.4	8.5	3
21	1.25	0.5	0.2	26.4	9	3
22	1.25	0.3	0.4	26.4	10	5.8
23	1.25	0.2	0.5	26.4	8.5	5.8
24	1.25	0.2	0.5	26.4	9	5.8
25	1.25	0.2	0.5	26.4	10	5.8
26	1.25	0.4	0.3	26.4	9.5	10
27	1.25	0.4	0.3	26.4	10	10

Table 4 (continued)

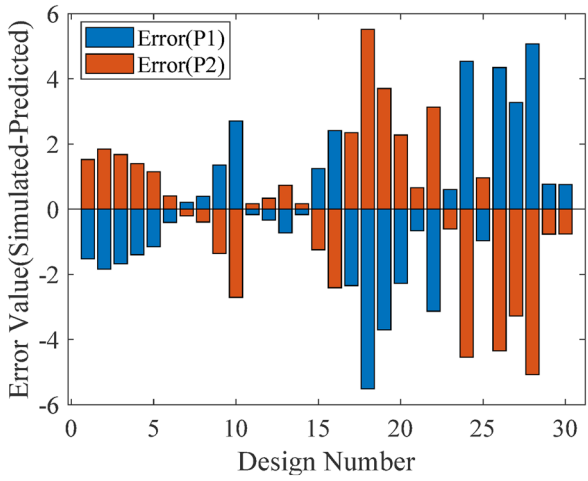
Design Number	Input Properties					
	Width (μm)	h_1 (μm)	h_2 (μm)	L (μm)	W (μm)	L_1 (μm)
28	1.25	0.3	0.4	26.4	8	10
29	1.25	0.2	0.5	26.4	9	10
30	1.25	0.2	0.5	26.4	9.5	10
Output Variables						
Simulated value of P_1:P_2				Predicted value of P_1:P_2		
P_1		P_2		P_1		P_2
50		50		48.48		51.52
50		50		48.16		51.84
49.99		50		48.32		51.68
49.99		50		48.60		51.40
49.99		50		48.85		51.15
49.99		50		49.60		50.40
49.99		50		50.20		49.80
49.99		50		50.40		49.60
50.00		49.99		51.36		48.64
53.84		46.16		56.55		43.45
49.80		50.20		49.63		50.37
49.84		50.16		49.51		50.49
50.13		49.87		49.39		50.61
49.64		50.36		49.48		50.52
48.49		51.51		49.73		50.27
53.68		46.32		56.09		43.91
59.30		40.70		56.96		43.04
61.48		38.52		55.97		44.03
60.08		39.92		56.38		43.62
58.87		41.13		56.60		43.40
58.58		41.42		57.92		42.08

Table 4 (continued)

Output Variables			
Simulated value of P_1:P_2		Predicted value of P_1:P_2	
P_1	P_2	P_1	P_2
67.76	32.24	64.63	35.37
65.92	34.08	66.53	33.47
69.27	30.73	73.82	26.18
72.96	27.04	72.00	28.00
80.56	19.44	84.91	15.09
79.86	20.14	83.14	16.86
75.42	24.58	80.50	19.50
89.49	10.51	90.26	9.74
91.73	8.27	92.50	7.50



(a)



(b)

Fig. 5 Test output **a** real and predicted results **b** error results

Such a strong effect clearly demonstrates that L_1 should be prioritized in the optimization process. In conclusion, the analyses in Fig. 6 show that each input parameter affects the outputs differently, with some exhibiting linear positive or negative effects while others display nonlinear behavior. These results underline the originality of the study. The proposed DNN model not only achieved high prediction accuracy but also quantified the influence of input variables on the outputs, providing a practical framework for design optimization.

The physical mechanism underlying the dominant influence of L_1 can be understood by examining its geometric role within the MMI coupling region. As illustrated in Fig. 2c, L_1 defines the length of the rectangular section removed from the upper-left corner of the multimode region. When $L_1 = 0$, the structure retains full lateral symmetry and the self-imaging condition of MMI theory produces a balanced 50:50 power distribution at the output ports. As L_1 increases, the effective boundary condition along the upper edge of the multi-

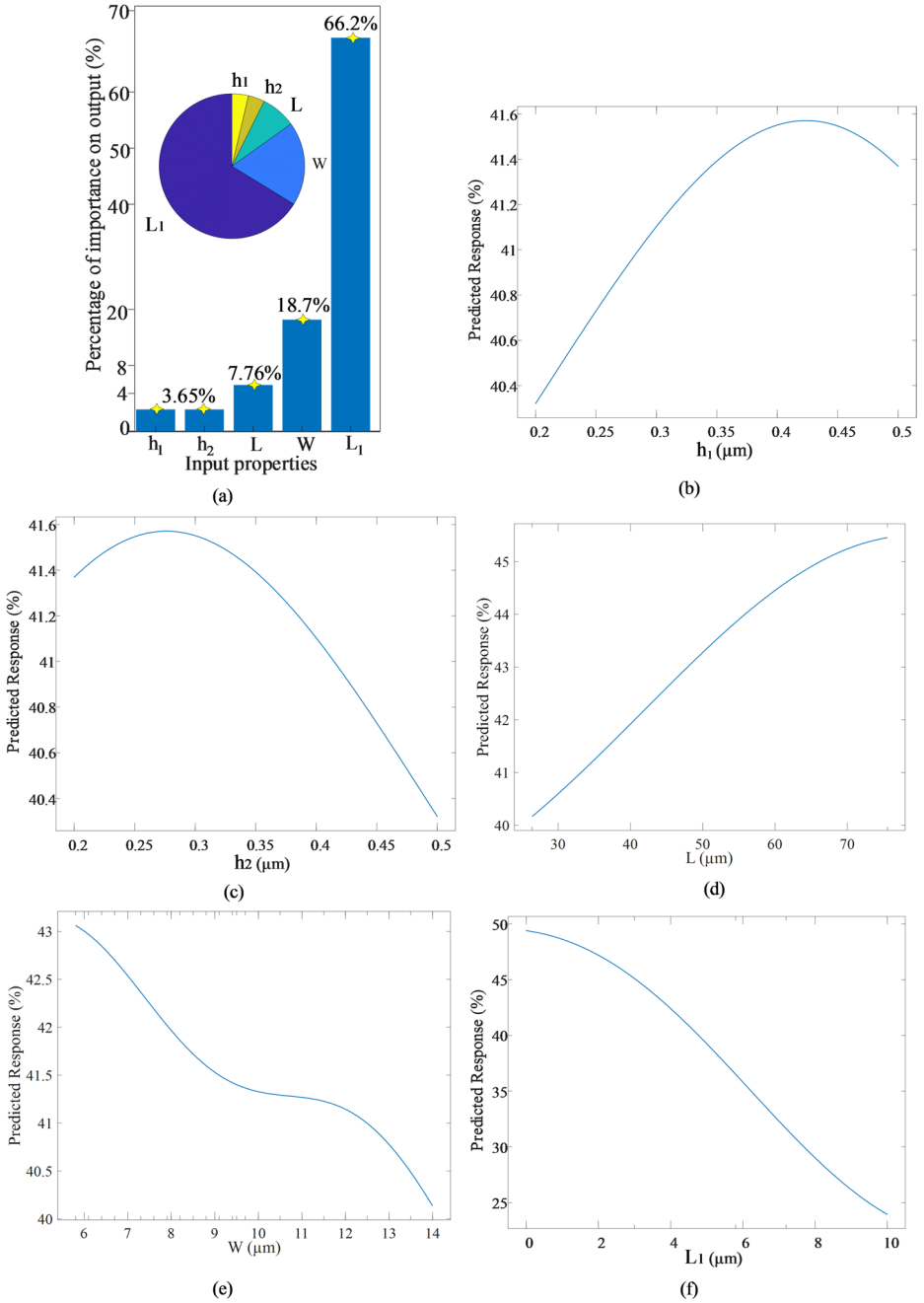


Fig. 6 Comparative analysis of the effects of input parameters on the model output: **a** importance distribution **b** h_1 **c** h_2 **d** L **e** W **f** L_1

mode region is progressively modified, breaking the lateral symmetry of the optical field. This geometric asymmetry introduces a differential propagation constant Δβ between the modes guided toward the upper and lower output ports. According to coupled-mode theory, the accumulated phase difference between these modes evolves as $\Delta\varphi = \Delta\beta \cdot L_1$, meaning that even a modest increase in L_1 produces a disproportionately large phase imbalance. This accumulated Δφ directly governs the output power ratio P_1/P_2, which explains why L_1 emerges as the single most influential parameter (~66.2%) in the sensitivity analysis of Fig. 6a. The progressive field asymmetry corresponding to increasing L_1 values is also directly observable in the EME simulation results of Fig. 3(a–d), where the output intensity distributions transition from near-symmetric (50:50) to strongly unbalanced (80:20) configurations as L_1 grows. The quantitative influence of W on the excitation efficiency of higher-order modes is captured by the sensitivity curve shown in Fig. 6e. Within the framework of MMI theory, an increase in W broadens the transverse modal spectrum of the multimode region, facilitating the excitation of higher-order modes and reducing the mode overlap integral between the fundamental input mode and the guided output modes. This progressive disruption of the self-imaging condition degrades the uniformity of power distribution at the output ports, which is directly reflected in the strong negative sensitivity of the output ratio to W (~18.7%). The monotonically decreasing profile of the Fig. 6e curve is therefore not merely a statistical trend but a quantitative signature of this mode-excitation mechanism operating within the multimode coupling region.

In summary, these findings demonstrate that the DNN model does not merely capture statistical correlations but effectively reflects the underlying electromagnetic coupling, mode overlap, and interference-length mechanisms governing the splitter operation. By quantitatively identifying the relative contributions of each design variable and physically explaining the dominant roles of L_1 and W, the proposed approach provides a reliable and interpretable framework for guiding the optimization of LNOI-based power splitters. This physics-consistent insight significantly reduces the reliance on extensive electromagnetic simulations and enables efficient, data-driven design-space exploration.

3.3 Integrated DNN–optimization framework for forward prediction and inverse design

A MATLAB-based application was developed to provide an integrated computational framework that performs both forward prediction (computing the output power ratio from given geometric parameters) and inverse design (determining the optimal geometric parameters corresponding to a target output power ratio). The system is built upon a dataset generated through Lumerical EME/FDTD simulations, a DNN-based regression model trained on this dataset, and an accompanying optimization module responsible for solving the inverse problem. The operation and graphical interface of the system are illustrated in Fig. 7(a)–7(d).

Figure 7(a) illustrates the inverse design interface, in which the user inputs only the desired P_1:P_2 output power ratio (e.g., 50:50) in the format "50:50." After normalizing the target ratio, the system jointly employs the pre-trained DNN-based forward prediction model and a gradient-based optimization algorithm. The optimization process is carried out by minimizing the discrepancy between the target power ratio and the DNN-predicted output ratio, with the learning rate set to 10^{-2} and the maximum number of iterations limited to 5000. Convergence is assumed when the change in the error between successive iterations

Fig. 7 MATLAB-based DNN optimization app showing **a** inverse design interface, **b** forward prediction interface, **c** inverse design example for the 50:50 target, and **d** forward prediction result using (**c**)

falls below 10^{-6} or when the maximum iteration count is reached. Throughout this process, the geometric parameters Width, h_1, h_2, L, W, and L_1 are constrained within physically feasible ranges determined by LNOI fabrication tolerances and the validity limits of the EME simulations. As a result, the algorithm automatically computes the optimal set of geometric parameters corresponding to the user-defined power splitting ratio.

Figure 7(b) shows the forward prediction interface, where the user manually enters six geometric parameters. In this module, the system solely employs the trained DNN regression model to rapidly estimate the resulting P_1:P_2 output ratio, without invoking any optimization procedure. Similar to Fig. 7(a), no example values are presented here; the figure illustrates the core forward prediction interface utilized for evaluating design candidates.

Figure 7(c) illustrates an example generated by the inverse design module. In this demonstration, the user specified a target ratio of 50:50, and the integrated DNN optimization

framework produced a feasible solution. For this target, the system computed approximately Width = 1.092, h_1 = 0.259, h_2 = 0.438, L = 32.027, W = 7.907, and L_1 = 4.801. This figure illustrates how the inverse design process synthesizes a set of geometric parameters that closely satisfy the specified output power ratio.

Figure 7(d) presents an example of the forward prediction module. In this scenario, the user entered the same geometric parameters obtained from the inverse design module into the forward prediction interface. The DNN model then evaluated these inputs and yielded an output power ratio of approximately P1:P2 = 51.89: 48.11. This result demonstrates that the parameters computed in Fig. 3(c) indeed produce an output ratio very close to the target 50:50, confirming the consistency and reliability of the combined forward and inverse design workflow.

Figure 7(a) and Fig. 7(c) correspond to the optimization-assisted inverse design procedure, whereas Fig. 7(b) and Fig. 7(d) represent the purely DNN-based forward prediction method. This integrated approach enables the user to efficiently obtain both the optimal geometric design variables for a desired output power ratio and the performance estimation of a given set of parameters. Consequently, the developed MATLAB application serves as an effective, fast, and high-accuracy computational tool for designing and evaluating the performance of LNOI-based power splitters.

To further evaluate the fabrication robustness of the inverse design output, a process tolerance analysis was performed on the optimal parameter set obtained for the 30:70 target ratio. Each geometric parameter was perturbed by +0.05 μm to simulate a worst-case fabrication scenario in which all parameters simultaneously deviate in the same direction by the maximum expected process error, reflecting typical LNOI etching and thin-film deposition tolerances. The perturbed parameter set was then evaluated using the trained DNN forward prediction model. As shown in Fig. 8, the original inverse design result (Fig. 8(a)) yields Width = 0.811, h_1 = 0.330, h_2 = 0.393, L = 39.237, W = 6.441, and L_1 = 0.223 μm for the 30:70 target. When all six parameters are simultaneously increased by +0.05 μm and fed into the forward prediction module (Fig. 8(b)), the predicted output ratio is P_1:P_2 = 30.56:69.44, corresponding to a deviation of only ± 0.56% from the target 30:70 ratio. This value is substantially below the ± 1.5% fabrication compatibility threshold and represents a worst-case bound, since in practice fabrication errors across different parameters are statistically independent and do not all shift simultaneously in the same direction. These results confirm that the inverse design module produces solutions that are inherently robust to the dimensional tolerances of practical LNOI fabrication processes.

The operation efficiency of the developed tool was evaluated on a desktop workstation equipped with an AMD Ryzen 9 7900 processor, 32 GB RAM, and MATLAB R2024b. Under these conditions, a single forward prediction corresponding to one DNN inference pass through the 16,065-parameter network completes in approximately 1–3 milliseconds. A complete inverse design run comprising 5,000 gradient-based optimization iterations requires approximately 3–8 s in total, depending on convergence behavior; in practice the optimization often terminates well before the maximum iteration count is reached, since convergence is assumed when the per-iteration error falls below 10^{-6}, further reducing the average run time. In comparison, a single Lumerical EME simulation of the same LNOI power splitter geometry under the mesh and boundary condition settings employed in this study requires approximately 9–11 min on the same workstation. This corresponds to a computational speed-up factor of approximately 180,000–660,000 for a single forward pre-

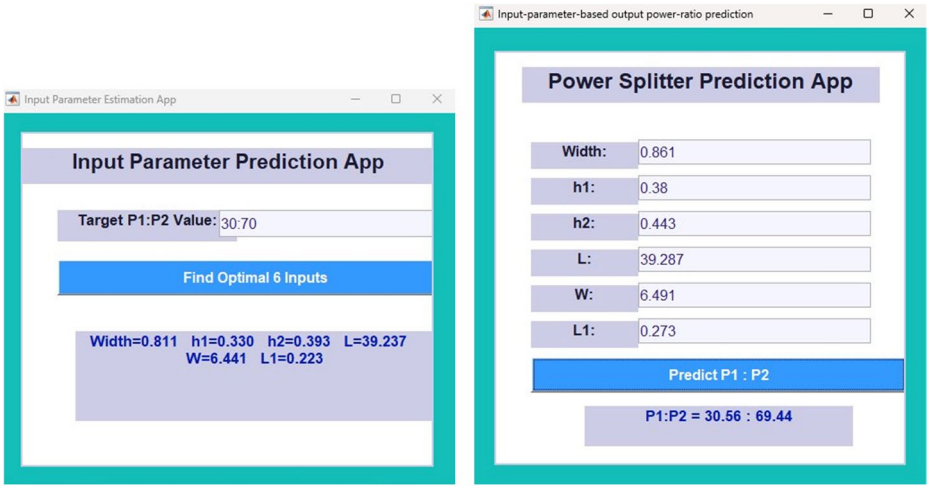


Fig. 8 Process tolerance robustness analysis: **a** inverse design result for the 30:70 target ratio (Width = 0.811, h_1 = 0.330, h_2 = 0.393, L = 39.237, W = 6.441, L_1 = 0.223 μm), and **b** forward prediction result after applying a + 0.05 μm perturbation to all six parameters simultaneously, yielding P_1:P_2 = 30.56:69.44 and confirming a worst-case power ratio deviation of only ± 0.56%, well within the ± 1.5% fabrication compatibility threshold

diction and 68–220 for a complete inverse design cycle relative to a single EME simulation. These results demonstrate that the proposed DNN-based framework reduces design evaluation time from the minute scale to the millisecond scale for forward prediction, and from multiple sequential simulation runs each requiring nearly ten minutes to a single-digit second timescale for inverse design, confirming the substantial engineering practicality of the developed MATLAB application for real-world photonic device design tasks.

3.4 Comparative validation on literature-based data

To evaluate the generalisation ability of the proposed DNN-based model, the trained network was assessed using an external dataset reported in the literature. The results were obtained for the P_1 output during the evaluation. The reason for this is that P_1 and P_2 are complementary values. The input-output pairs listed in Table 5 were used to compare the model's prediction accuracy on independent data. The results confirm that the model demonstrated robustness and transferability, showing consistent and satisfactory performance not only on its own training data but also on various data combinations in the literature.

In Fig. 9a, the results of a total of twelve designs corresponding to two different literature references (six designs for each reference) are presented. The graph compares the power distribution ratios P_1 and P_2 obtained from simulations (real values) and from the DNN model predictions for each design. The first reference, shown in the gray-shaded region on the left, corresponds to the study (Li et al. 2024), while the second reference, shown in the green-shaded region on the right, corresponds to the study (Lin et al. 2023). In both reference sets, the model's prediction curves exhibit a high degree of similarity to the actual data.

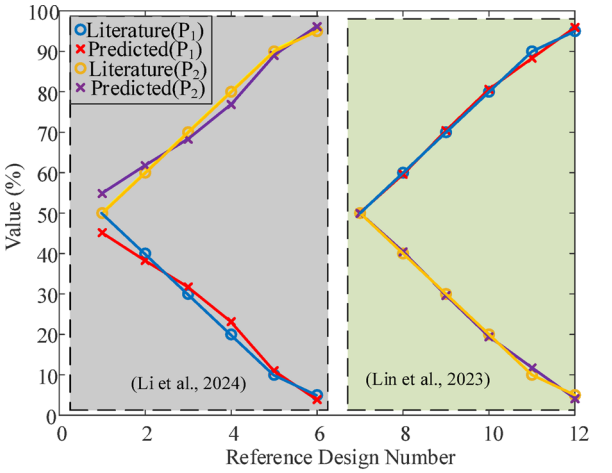
Table 5 Literature data and results for DNN-based design

Input Variables							
Reference	Reference Design Number	Width (μm)	h_1 (μm)	h_2 (μm)	L (μm)	W (μm)	L_1 (μm)
(Li et al. 2024)	1	1.5	0.3	0.3	26.4	5.8	0
(Li et al. 2024)	2	1.5	0.3	0.3	27.9	5.8	3.1
(Li et al. 2024)	3	1.5	0.3	0.3	29.2	5.8	5.5
(Li et al. 2024)	4	1.5	0.3	0.3	31	5.8	7.9
(Li et al. 2024)	5	1.5	0.3	0.3	32.9	5.8	12.5
(Li et al. 2024)	6	1.5	0.3	0.3	33.3	5.8	16.4
(Lin et al. 2023)	7	0.8	0.18	0.18	57.5	9.6	0
(Lin et al. 2023)	8	0.8	0.18	0.18	62	9.6	17
(Lin et al. 2023)	9	0.8	0.18	0.18	75	9.6	25
(Lin et al. 2023)	10	0.8	0.18	0.18	82	9.6	35
(Lin et al. 2023)	11	0.8	0.18	0.18	103	9.6	66
(Lin et al. 2023)	12	0.8	0.18	0.18	108	9.6	80

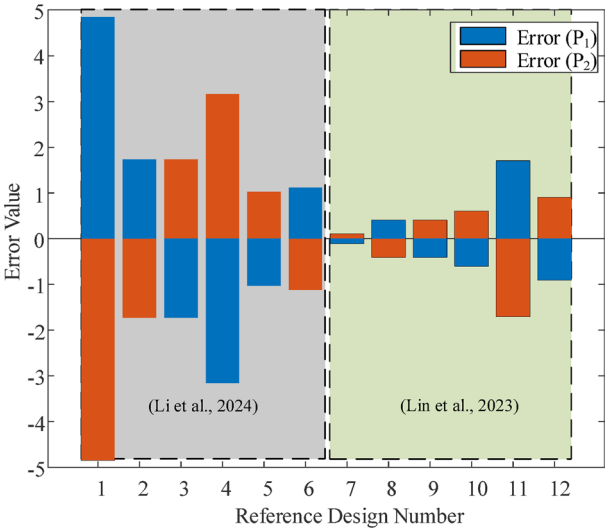
Output Variable			
Literature value of P_1:P_2		Predicted value of P_1:P_2	
P_1	P_2	P_1	P_2
50	50	45.16	54.84
40	60	38.27	61.73
30	70	31.73	68.27
20	80	23.16	76.84
10	90	11.03	88.97
5	95	3.88	96.12
50	50	50.1	49.9
60	40	59.6	40.4
70	30	70.4	29.6
80	20	80.6	19.4

Table 5 (continued)

Output Variable			
Literature value of P_1:P_2		Predicted value of P_1:P_2	
P_1	P_2	P_1	P_2
90	10	88.3	11.7
95	5	95.9	4.1



(a)



(b)

Fig. 9 Literature output **a** real and predicted results, **b** error results

Notably, the P_1 and P_2 values vary complementarily, indicating that the DNN model has successfully learned to strike a balance between the two output ports.

In Fig. 9b, the error values for the same designs are illustrated. The bar charts represent the differences between the real and predicted results (Real − Predicted) in percentage terms. The error values remain generally very low, confirming the high prediction performance of the model. In the initial reference group (left part), few designs demonstrate variances of approximately ± 4%, yet the overarching pattern remains intact; in the subsequent reference group (right section), the errors are much less significant, nearing zero. This outcome indicates that the model is capable of generating accurate predictions even when tested with datasets obtained from different literature sources. The validation metrics further substanti-

ate the model's performance, yielding $R = 0.991$, RMSE = 1.98, and MAPE = 3.42%. These results verify that the proposed DNN architecture achieves high predictive accuracy and low error rates, even when applied to independent datasets drawn from the literature.

In summary; Fig. 9 demonstrates that the proposed DNN-based model exhibits not only high accuracy but also strong generalization capability, extending beyond the training dataset to independent datasets reported in the literature. The close agreement between real and predicted values confirms the model's physical consistency, while the low error percentages verify its superior predictive performance even in literature-based validation scenarios.

4 Conclusion

In this study, a novel DNN-based framework was proposed and successfully implemented for predicting and analyzing the performance of arbitrary-ratio power splitters on the LNOI platform. The results demonstrated that conventional machine learning algorithms such as SVM, GPR, and Tree-based models were insufficient to capture the complex and nonlinear dependencies between geometric design parameters and optical responses. In contrast, the proposed DNN architecture exhibited superior predictive accuracy, achieving high correlation coefficients (R^2 = 0.95–0.97) and remarkably low RMSE and MAPE values across the training, validation, and testing phases. The model, optimized using the Adam algorithm with a batch size of 64 and 5000 training epochs, provided the best balance between accuracy, convergence stability, and generalization capability. Beyond delivering highly accurate predictions of output power ratios, the DNN framework enabled a quantitative assessment of the relative importance of each geometric variable, revealing that parameters such as L1 and W played dominant roles in shaping device behavior. Furthermore, validation on independent datasets from the literature confirmed the robustness and transferability of the model, demonstrating excellent agreement with previously reported experimental findings. An additional contribution of this work is the development of an interactive MATLAB application that integrates both the forward prediction and inverse design capabilities of the DNN framework. This tool allows users to estimate output power ratios directly from user-defined geometric parameters and compute optimal parameter sets corresponding to a target output ratio through a combined DNN–optimization workflow. The app significantly streamlines device evaluation, enables rapid design-space exploration, and provides a practical interface for researchers engaged in photonic device design.

In conclusion, the proposed approach substantially simplifies and accelerates the design process of LNOI-based power splitters by eliminating the need for computationally expensive electromagnetic simulations. The DNN framework reduces design complexity, improves interpretability of parameter–performance relationships, and lays the groundwork for future advancements in optimization, inverse design, and fast prototyping of integrated photonic components. The outcomes of this research highlight the potential of data-driven methodologies to transform the design paradigms of next-generation photonic integrated circuits.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11082-026-08835-y.

Acknowledgements This work has been supported by the Scientific Research Project Coordination Unit (BAP) of Bursa Uludag University, Turkey, with Project Numbers FGA-2024-1745 and FAY-2024-1777.

Author contributions Huriye Gencal: Formal Analysis, Methodology, Investigation, Validation, Writing – Original Draft.Abdullah Aksoy: Methodology, Software, Investigation, Visualization, Writing – Original Draft.Enes Yiğit: Supervision, Review & Editing, Visualization.Umut Aydemir: Conceptualization, Supervision, Writing – Review & Editing.Mustafa Demirtaş: Conceptualization, Supervision, Data Curation, Visualization, Writing – Review & Editing.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability Code and processed datasets are publicly available at: https://github.com/aksoy1993/power-splitter.git.

Declarations

Competing interests The authors declare no competing interests.

References

Anderson, S.P., Webster, M.: Silicon Photonic Polarization-Multiplexing Nanotaper for Chip-to-Fiber Coupling. J. Lightwave Technol. **34**(2), 372–378 (2016). https://doi.org/10.1109/JLT.2015.2498528

Asano, T., Noda, S.: Optimization of photonic crystal nanocavities based on deep learning. Opt. Express. **26**(25), 32704 (2018). https://doi.org/10.1364/oe.26.032704

Boes, A., Chang, L., Langrock, C., Yu, M., Zhang, M., Lin, Q., Lončar, M., Fejer, M., Bowers, J., Mitchell, A.: Lithium niobate photonics: Unlocking the electromagnetic spectrum. Science. **379**(6627), eabj4396 (2023). https://doi.org/10.1126/science.abj4396

Bogaerts, W., de Heyn, P., van Vaerenbergh, T., de Vos, K., Kumar Selvaraja, S., Claes, T., Dumon, P., Bienstman, P., van Thourhout, D., Baets, R.: Silicon microring resonators. Laser Photonics Reviews. **6**(1), 47–73 (2012). https://doi.org/10.1002/lpor.201100017

Bogaerts, W., Pérez, D., Capmany, J., Miller, D.A.B., Poon, J., Englund, D., Morichetti, F., Melloni, A.: Programmable photonic circuits. Nature. **586**(7828), 207–216 (2020). https://doi.org/10.1038/s41586-020-2764-0

Chung, K.K., Chan, H.P., Chu, P.L.: A 1 × 4 polarization and wavelength independent optical power splitter based on a novel wide-angle low-loss Y-junction. Opt. Commun. **267**(2), 367–372 (2006). https://doi.org/10.1016/j.optcom.2006.06.048

Dai, D., Liu, L., Gao, S., Xu, D.X., He, S.: Polarization management for silicon photonic integrated circuits. Laser Photonics Reviews. **7**(3), 303–328 (2013). https://doi.org/10.1002/lpor.201200023

Demirtaş, M., Özden, A., Açikbaş, E. et al. Extensive mode mapping and novel polarization filter design for ALD grown Al2O3 ridge waveguides. Opt. Quant. Electron. **48**, 357 (2016). https://doi.org/10.1007/s11082-016-0629-4

Deng, Q., Liu, L., Li, X., Zhou, Z.: Arbitrary-ratio 1x 2 power splitter based on asymmetric multimode interference. Opt. Lett. **39**(19), 5590–5593 (2014). https://doi.org/10.1364/OL.39.005590

Feng, H., Ge, T., Guo, X., Wang, B., Zhang, Y., Chen, Z., Zhu, S., Zhang, K., Sun, W., Huang, C., Yuan, Y., Wang, C.: Integrated lithium niobate microwave photonic processing engine. Nature. **627**(8002), 80–87 (2024). https://doi.org/10.1038/s41586-024-07078-9

Formisano, A., Tucci, M.: Machine Learning Approaches for Inverse Problems and Optimal Design in Electromagnetism. Electron. (Switzerland). **13**(7), 1–13 (2024). https://doi.org/10.3390/electronics13071167

Frellsen, L.F., Ding, Y., Sigmund, O., Frandsen, L.H.: Topology optimized mode multiplexing in silicon-on-insulator photonic wire waveguides. Opt. Express. **24**(15), 16866 (2016). https://doi.org/10.1364/oe.24.016866

Gencal, H., Demirtaş, M., Aydemir, U.: Lithium Niobate-Based Single-Mode Rib Waveguide: Design and Parameter Optimization. *2025 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 1–4. (2025). https://doi.org/10.1109/ISAS66241.2025.11101980

Hang, M.I.A.N.Z., Ang, C.H.W.: Monolithic ultra-high- Q lithium niobate microring resonator. Optica. **4**(12), 1536–1537 (2017). https://doi.org/10.1364/OPTICA.4.001536

Jalali, B., Fathpour, S.: Silicon photonics. In *Journal of Lightwave Technology* (Vol. 24, Issue 12, pp. 4600–4615). Institute of Electrical and Electronics Engineers Inc. (2006). https://doi.org/10.1109/JLT.2006.885782

Kang, C., Park, C., Lee, M., Kang, J., Jang, M.S., Chung, H.: Large-scale photonic inverse design: computational challenges and breakthroughs. In: Nanophotonics, vol. 13, pp. 3765–3792. Walter de Gruyter GmbH (2024). https://doi.org/10.1515/nanoph-2024-0127

Kong, A., Tobing, L.Y.M., Zhang, Y.: Photonic power splitter design based on deep learning and gradient descent method. In A. Adibi, S.-Y. Lin, & A. Scherer (Eds.), *Photonic and Phononic Properties of Engineered Nanostructures XIV* (Vol. 12896, p. 1289605). SPIE. (2024). https://doi.org/10.1117/12.3002496

Li, D., Li, J., Li, R., Liu, J.: The Design and Characterization of an Ultra-Compact Asymmetrical Multimode Interference Splitter on Lithium Niobate Thin Film. Photonics. **11**(1) (2024). https://doi.org/10.3390/photonics11010060

Liao, J., Tian, Y., Zhang, X., An, Y., Kang, Z.: Arbitrary ratio power splitter based on shape optimization for dual-band operation. Opt. Laser Technol. **172**(September 2023), 110495 (2024). https://doi.org/10.1016/j.optlastec.2023.110495

Lin, Y., Ke, W., Ma, R., Huang, F., Tan, H., Xu, J., Lin, Z., Cai, X.: Arbitrary-ratio 1 × 2 optical power splitter based on thin-film lithium niobate. Opt. Express. **31**(17), 27266 (2023). https://doi.org/10.1364/oe.497887

Liu, D., Tan, Y., Khoram, E., Yu, Z.: Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. ACS Photonics. **5**(4), 1365–1369 (2018). https://doi.org/10.1021/acsphotonics.7b01377

Liu, X., Sheng, Z., Zhao, Y., Gan, F.: Ultra-broadband on-chip power splitters for arbitrary ratios on silicon-on-insulator. *Optics Express*, *32*(2), 2029. (2024). https://doi.org/10.1364/oe.508058

Liu, Y., Kang, Z., Xu, H., Zhong, G., Zhang, R., Fu, C., Tian, Y.: Inverse Design of Multi-Port Power Splitter with Arbitrary Ratio Based on Shape Optimization. Nanomaterials. **15**(5), 1–11 (2025). https://doi.org/10.3390/nano15050393

Lyu, J., Kong, W., Pi, Y., Chen, Z., Xu, K., Wang, L., Yu, S.: Inverse-Designed 1x2 power splitter on X-Cut Thin-Film Lithium Niobate Platform by DUV Photonic Integration. Opt. Express. **33**(16), 34727–34735 (2025). https://doi.org/10.1364/oe.567729

Ma, W., Liu, Z., Kudyshev, Z.A., Boltasseva, A., Cai, W., Liu, Y.: Deep learning for the design of photonic structures. Nat. Photonics. **15**(2), 77–90 (2021). https://doi.org/10.1038/s41566-020-0685-y

Malkiel, I., Mrejen, M., Nagler, A., Arieli, U., Wolf, L., Suchowski, H.: Plasmonic nanostructure design and characterization via Deep Learning. Light: Sci. Appl. **7**(1) (2018). https://doi.org/10.1038/s41377-018-0060-7

Molesky, S., Lin, Z., Piggott, A.Y., Jin, W., Vucković, J., Rodriguez, A.W.: Inverse design in nanophotonics. Nat. Photonics. **12**(11), 659–670 (2018). https://doi.org/10.1038/s41566-018-0246-9

Novick, A., James, A., Dai, L.Y., Wu, Z., Rizzo, A., Wang, S., Wang, Y., Hattink, M., Gopal, V., Jang, K., Parsons, R., Bergman, K.: High-bandwidth density silicon photonic resonators for energy-efficient optical interconnects. Appl. Phys. Reviews. **10**(4) (2023). https://doi.org/10.1063/5.0160441

Pan, Z., Pan, X.: Deep Learning and Adjoint Method Accelerated Inverse Design in Photonics: A Review. Photonics. **10**(7) (2023). https://doi.org/10.3390/photonics10070852

Peurifoy, J., Shen, Y., Jing, L., Yang, Y., Cano-Renteria, F., DeLacy, B.G., Joannopoulos, J.D., Tegmark, M., Soljačić, M.: Nanophotonic particle simulation and inverse design using artificial neural networks. Sci. Adv. **4**(6), 1–7 (2018a). https://doi.org/10.1126/sciadv.aar4206

Peurifoy, J., Shen, Y., Jing, L., Yang, Y., Cano-Renteria, F., DeLacy, B.G., Joannopoulos, J.D., Tegmark, M., Soljačić, M.: Nanophotonic particle simulation and inverse design using artificial neural networks. Sci. Adv. **4**(6), 1–7 (2018b). https://doi.org/10.1126/sciadv.aar4206

Poberaj, G., Hu, H., Sohler, W., Günter, P.: Lithium niobate on insulator (LNOI) for micro-photonic devices. Laser \& Photonics Reviews. **6**(4), 488–503 (2012). https://doi.org/10.1002/lpor.201100035

Rangaraj, M., Minakata, M., Kawakami, S.: Low Loss Integrated Optical Y-Branch. J. Lightwave Technol. **7**(5), 753–758 (1989). https://doi.org/10.1109/50.19110

Reed, G.T., Mashanovich, G., Gardes, F.Y., Thomson, D.J.: Silicon optical modulators. In *Nature Photonics* (Vol. 4, Issue 8, pp. 518–526). (2010). https://doi.org/10.1038/nphoton.2010.179

Shen, Y., Harris, N.C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., Soljacic, M.: Deep learning with coherent nanophotonic circuits. Nat. Photonics. **11**(7), 441–446 (2017). https://doi.org/10.1038/nphoton.2017.93

Shen, R., Hong, B., Ren, X., Yang, F., Chu, W., Cai, H., Huang, W.: Recent progress on inverse design for integrated photonic devices: methodology and applications. J. Nanophotonics. **18**(01) (2024). https://doi.org/10.1117/1.jnp.18.010901

Soldano, L.B., Pennings, E.C.M.: Optical Multi-Mode Interference Devices Based on Self-Imaging: Principles and Applications. J. Lightwave Technol. **13**(4), 615–627 (1995). https://doi.org/10.1109/50.372474

Song, C., Zhang, Z., Wang, H., Chen, J., Zhang, K., Li, L., Lin, T., Chen, S., Lu, J., Ni, Z.: Ultracompact and broadband Si3N4 Y-branch splitter using an inverse design method. Opt. Express. **32**(26), 46080 (2024). https://doi.org/10.1364/oe.542341

Soref, R.: The past, present, and future of silicon photonics. IEEE J. Sel. Top. Quantum Electron. **12**(6), 1678–1687 (2006). https://doi.org/10.1109/JSTQE.2006.883151

Soref, R.: Silicon photonics: A review of recent literature. Silicon. **2**(1), 1–6 (2010). https://doi.org/10.1007/s12633-010-9034-y

Takahashi, K., Nonaka, S.: Optical Directional Coupler. Opt. InfoBase Conf. Papers. **Part F11803**, 2004–2006 (1977)

Vivien, L., Pavesi, L.: Handbook of silicon photonics. Taylor & Francis (2016)

Wang, C., Zhang, M., Chen, X., Bertrand, M., Shams-Ansari, A., Chandrasekhar, S., Winzer, P., Lončar, M.: Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. Nature. **562**(7725), 101–104 (2018). https://doi.org/10.1038/s41586-018-0551-y

Xu, Q., Schmidt, B., Pradhan, S., Lipson, M.: Micrometre-scale silicon electro-optic modulator. Nature. **435**(7040), 325–327 (2005). https://doi.org/10.1038/nature03569

Yariv, A.: *Coupled-Mode Theory for Guided-Wave Optics*. *9*. (1973)

Yigit, E., Hayber, Ş.E., Aydemir, U.: ANN-based estimation of MEMS diaphragm response: An application for three leaf clover diaphragm based Fabry-Perot interferometer. Measurement. **199**, 111534 (2022). https://doi.org/10.1016/j.measurement.2022.111534

Yousefi, P., Khalid, M., Petruzzelli, V., Calò, G.: Design of thin-film lithium niobate power splitters and combiners based on multimode interference. Opt. Quant. Electron. **57**(3), 1–23 (2025). https://doi.org/10.1007/s11082-025-08060-z

Zhu, D., Shao, L., Yu, M., Cheng, R., Desiatov, B., Xin, C.J., Hu, Y., Holzgrafe, J., Ghosh, S., Shams-Ansari, A., Puma, E., Sinclair, N., Reimer, C., Zhang, M., Lončar, M.: Integrated photonics on thin-film lithium niobate. Adv. Opt. Photon. **13**(2), 242–352 (2021). https://doi.org/10.1364/AOP.411024

Demirtaş, M., Özden, A., Açikbaş, E. et al. Extensive mode mapping and novel polarization filter design for ALD grown Al2O3 ridge waveguides. Opt Quant Electron 48, 357 (2016) https://doi.org/10.1007/s11082-016-0629-4

Authors and Affiliations

Huriye Gencal[1,2] · Abdullah Aksoy[2] · Enes Yigit[2,3] · Umut Aydemir[2,3] · Mustafa Demirtas[2,3]

✉ Mustafa Demirtas
mustafademirtas@uludag.edu.tr

[1] Department of Electronic Communication Technology, Mudanya University, Bursa 16940, Türkiye

[2] Department of Electrical and Electronics Engineering, Bursa Uludag University, Bursa 16059, Türkiye

[3] Department of Optics and Photonics Engineering, Bursa Uludag University, Bursa 16059, Türkiye