NUCL SCI TECH (2022) 33:94
https://doi.org/10.1007/s41365-022-01069-z

A non-invasive diagnostic method of cavity detuning based on a convolutional neural network

Liu-Yuan Zhou[1,2] · Hao Zha[1,2] · Jia-Ru Shi[1,2] · Jia-Qi Qiu[1,2] · Chuan-Jing Wang[1,2] · Yun-Sheng Han[1,2] · Huai-Bi Chen[1,2]

Received: 7 April 2022 / Revised: 7 June 2022 / Accepted: 9 June 2022 / Published online: 20 July 2022

Abstract As modern accelerator technologies advance toward more compact sizes, conventional invasive diagnostic methods of cavity detuning introduce negligible interference in measurements and run the risk of harming structural surfaces. To overcome these difficulties, this study developed a non-invasive diagnostic method using knowledge of scattering parameters with a convolutional neural network and the interior point method. Meticulous construction and training of the neural network led to remarkable results on three typical acceleration structures: a 13-cell S-band standing-wave linac, a 12-cell X-band traveling-wave linac, and a 3-cell X-band RF gun. The trained networks significantly reduced the burden of the tuning process, freed researchers from tedious tuning tasks, and provided a new perspective for the tuning of side-coupling, semi-enclosed, and total-enclosed structures.

Keywords Cavity detuning · Convolutional neural network · Equivalent circuit

1 Introduction

Frequency detuning is an ineluctable concern for accelerator cavities. Cavity frequency detuning can be caused by a drift in temperature and humidity during machining, deformation of thermal stress during welding, vibration caused by mechanical movement during transportation, and breakdown of RF pulses during conditioning [1]. Concentrating solely on the machining process, the derivative of TM_{010} mode frequency of a cylindrical resonator with respect to its radius yields the following estimation formula: $\mathrm{d}f[\mathrm{MHz}] \approx -(2.9 \times 10^7)/r^2 \mathrm{d}r[\mu\mathrm{m}]$. For X-band structures, an error of 1 μm in the cavity radius results in frequency detuning of approximately 1 MHz.

An accelerator can be considered equivalent to a coupled-cavity chain. The field amplitude and phase advance of each cavity are determined by the frequencies and couplings of the chain. To guarantee proper field distributions and phase advancements, the accelerator cavities must be tuned prior to operation [2]. Two conventional methods generally work in low-frequency bands and even in some X-band structures. One is the so-called SLAC-type method that diagnoses cavity detuning by inserting two inflexible conductive probes into the accelerator and measuring the resonant frequencies of each individual cavity and neighbor-coupled cavities one by one [3]. To measure the frequency of each resonant cavity accurately, the position of the probes must be adjusted carefully so that the resonant peak signal of the vector network analyzer no longer drifts with small movements of the probe. However, this is a time-consuming task. In addition to the growing risk of harming the soft inner surface, the measurement results become sensitive to the probe positions for cavities in higher-frequency bands because of their more compact

This work was supported by the National Natural Science Foundation of China (No. 11922504).

✉ Jia-Ru Shi
shij@tsinghua.edu.cn

[1] Department of Engineering Physics, Tsinghua University, Beijing 100084, China

[2] Key Laboratory of Particle and Radiation Imaging, Tsinghua University, Beijing 100084, China

size. The second method, initially verified by Khabiboulline et al. in 1995, is to measure the field distributions by bead-pulling to invert the correlative cavity frequencies [4]. Later, Shi et al. extended this method to the coupler tuning process [5], and Yang et al. simplified the equivalent circuit and solved the matrix coefficients directly from field distributions [6]. The bead-pulling method is an on-axis field-distribution measurement technique that uses a tensioned string to pull a dielectric bead, and calculates the on-axis field amplitude by measuring the frequency shift caused by the field perturbation of the bead. However, applying the bead-pulling technique to non-through structures, such as RF guns or side-coupling structures, is difficult and imprecise. Moreover, the bead-pull line and rotations of the bead may cause considerable frequency drift in higher-frequency bands. Consequently, non-invasive diagnostic methods have attracted considerable attention. An emerging idea is based on the scattering parameters of the structures. Habel et al. were the first to compute the detuning of a superconducting linac using its dispersion parameters [7]. Owing to the lossless simplification caused by superconductivity, the equivalent circuit of the superconducting linac has the form of a tridiagonal matrix, which is easily solved using the conventional least-squares method. However, the situation is different at room temperature. Ni et al. devised a genetic algorithm to diagnose a normal conducting linac based on its scattering characteristics without the lossless simplification introduced by superconducting [8]. Complex numbers were added to the matrix. Ni et al. applied the nonlinear least-squares method with the NL2SOL operator to accelerate the convergence speed of the algorithm. Unfortunately, because of its strong reliance on the starting datasets, this method performed poorly on actual accelerators.

As a new discipline that is developing rapidly, artificial intelligence has shown increasing advantages in feature extraction and data modeling. At the junction of accelerators and artificial intelligence, groundbreaking work such as the computation of space charge force, beam line operation, and quench error diagnosis have been accomplished. Knowledge of Taylor maps was first reported in a polynomial neural network by Ivanov and Agapov for fast simulations of beam dynamics. The trained network approximated the dynamic beam system with perfect accuracy. With additional experimental data as the network input, they provided a means to tune network weights in real time [9]. Based on the same idea, Zou et al. successfully applied the asynchronous advantage actor–critic machine learning algorithm to the real-time tuning of the low-energy beam transport section (LEBT) of the Xi'an Proton Application Facility [10]. Considering the two-dimensional phase-space as a spatial image, Ren et al. simulated thousands of pairs of electron beams and X-ray power profiles to train convolutional neural networks, and used the trained networks to predict the X-ray power profiles with the input of electron beam phase spaces. This approach demonstrated a significant improvement over the traditional algorithm for a range of conditions [11]. Kain et al. used reinforcement learning algorithms that learned the optimal policy for a certain control problem and increased the efficiency of optimization algorithms for accelerator controls. They developed a continuous model-free reinforcement learning deep network with up to 16 degrees of freedom that can avoid the time-consuming exploration phase required for numerical optimizers after training [12]. With the knowledge of electromagnetic-based modeling, Rayas-Sánchez developed RF circuit designs using artificial neural networks. By designing the neural network training scheme to incorporate available knowledge, which can be obtained from an empirical equivalent circuit model based on quasi-static approximations, the knowledge-based neural network (KBNN) approach has become the most mature and automated technique for the development of neuromodels of RF circuits [13].

Following the footprints of pioneers, this study offers a novel diagnostic approach for cavity detuning based on convolutional neural networks (CNN) and the interior point method (IPM). The measured S_{11} is used as input to the CNN to obtain a coarse estimate of cavity detuning, and the IPM optimizes the coarse model based on the derivative computed from formulaic circuit theory. The remainder of this paper is organized as follows: Sect. 2 explains the diagnosis steps, describes the construction of the convolution neural network, and computes the derivatives of scattering parameters with respect to the frequency of each cavity; Sect. 3 documents the experimental results for three typical structures, including an X-band traveling-wave linac, an S-band standing-wave linac, and a 3-cell X-band RF gun, and discusses the performance of the method with respect to errors in other cavity parameters and sampling noise; finally, Sect. 4 concludes with a brief summary.

2 Method

Owing to the advantage of easy accessibility, the scattering parameters (e.g., S_{11}) are considered to be the starting point of our non-invasive diagnostic method. In addition, for most linacs with a completed package of electron guns and dose conversion targets, the scattering parameters are the only physical quantities accessible that reflect the RF states of the linac. The common finishing error is approximately 10 μm to 20 μm, resulting in a frequency detuning of approximately 10 to 20 MHz for X-band structures (11.424 MHz) or 5 to 10 MHz for S-band

structures (2.998 MHz); therefore, the diagnosis of cavity detuning is precisely a constrained optimization problem that involves finding a set of resonant frequencies ω_i in the above constrained zone to minimize $||S_{11,c} - S_{11,m}||^2$, where $S_{11,c}$ represents the S_{11} calculated by the diagnostic method, while $S_{11,m}$ is the measured S_{11} in reality. Although many traditional algorithms are available for solving optimization problems, the diagnosis of cavity detuning is strongly non-convex and nonlinear. Under these circumstances, traditional algorithms are highly likely to converge to a local optimum. Inspired by the applications of neural networks in the design of RF devices [14], the powerful fitting ability of artificial neural networks may provide a coarse estimate of cavity detuning close to the global optimum, and empower a traditional algorithm to cross the local optima in the constrained multidimensional space of the cavity diagnosis problem. Therefore, the diagnostic method for cavity detuning is divided into two steps. First, the scattering parameters measured from the waveguide of the structure coupler are normalized as the input of the neural network to obtain a coarse approximation of the cavity detuning. Starting from this coarse estimate, the IPM algorithm [15] based on an equivalent circuit model is applied to compute the fine detuning parameters. The role of the neural network is to circumvent the limitations of local optima in traditional optimization algorithms.

2.1 Construction of the neural network

The first step of the diagnostic method is to train a proper artificial neural network to predict the frequency detuning of each cavity based on the measured S_{11}. Hopfield neural networks (HNN), recurrent neural networks (RNN), and convolutional neural networks are three types of networks that are generally used in data regression prediction. Of these, convolutional neural networks (CNNs) have been widely used in the field of computer vision. For the diagnostic problem, the input of the network should be a vector whose elements correspond to the measured values S_{11} from the input port, whereas the output of the network is a vector whose elements correspond to the frequency of each cavity. As shown in Fig. 3, the S_{11} of an accelerator has significant characteristics that can be considered as a superposition of multiple Lorentzian resonance peaks, each of which contains information regarding the resonant frequency and Q-factor of the corresponding mode. Considering the S_{11} signals as a one-dimensional picture, many of the experiences learned from using convolutional neural networks for computer vision problems can be applied to our diagnostic problem. Therefore, we intuitively choose the CNN to process the input S_{11} information. We hope to obtain suitable convolution kernels through network training to express the intrinsic connection between the scattering parameters and detuning. After several cycles of modifications and debugging, the CNN, as shown in Fig. 1, comprises an input layer, a drop layer, a fully connected layer, an output layer with the mean square error as its loss function, and three convolutional units consisting of a convolution layer, normalization layer, rectified linear unit (ReLU), and pooling layer [16].

The S_{11} value of a normal conducting accelerator measured from its coupler is a complex array within a circle of radius 1. To take full advantage of the information in the real and imaginary parts of the measured S_{11}, the input layer of the CNN is a two-dimensional array composed of the magnitude and phase parts of the S_{11} measured from the feeding coupler. Each dimension contains measurements of 2048 frequency points. The frequency band of the measurement varies with the accelerator design. To avoid the difference in the weight sizes of the two dimensions of the input data, both the magnitude and phase parts of the input S_{11} are normalized to (0,1) using Eq. (1).

$$CNN_{\text{input}} = \begin{bmatrix} 1 - mag(S_{11}) \\ phase(S_{11})/\pi \end{bmatrix} \tag{1}$$

To accelerate the training and reduce the sensitivity to network initialization, the second layer is a standard batch normalization layer. This first calculates the mean and standard deviation of the input mini-batch and normalizes the input mini-batch by subtracting the mean and dividing by the standard deviation. The input mini-batch is then added and scaled by a learnable offset factor β and learnable scale factor γ.

The third layer is the first convolutional layer. This applies 48 sliding convolutional filters to the layer input, computes the dot product of the learnable weights of the filters with the layer input, and then adds a learnable bias term to the result. To better capture the resonance peak characteristics of the input, the size of the convolution kernel was first set to [64,1]. The length of the kernel is close to the full width at half maximum of the S_{11} resonant peaks of normal conducting accelerators. However, the large size of the kernel was found to slow the training speed and result in overfitting during numerical tests. Therefore, the kernel size was reset to [4,1]. With the following pooling process, the convolution kernel can still capture long-range characteristics.

Another batch normalization layer follows the convolution layer. The fifth layer is an activation layer, with the ReLU function as its activation function. The ReLU function performs a threshold operation on the input elements. Layer input values larger than 0 are set equal to

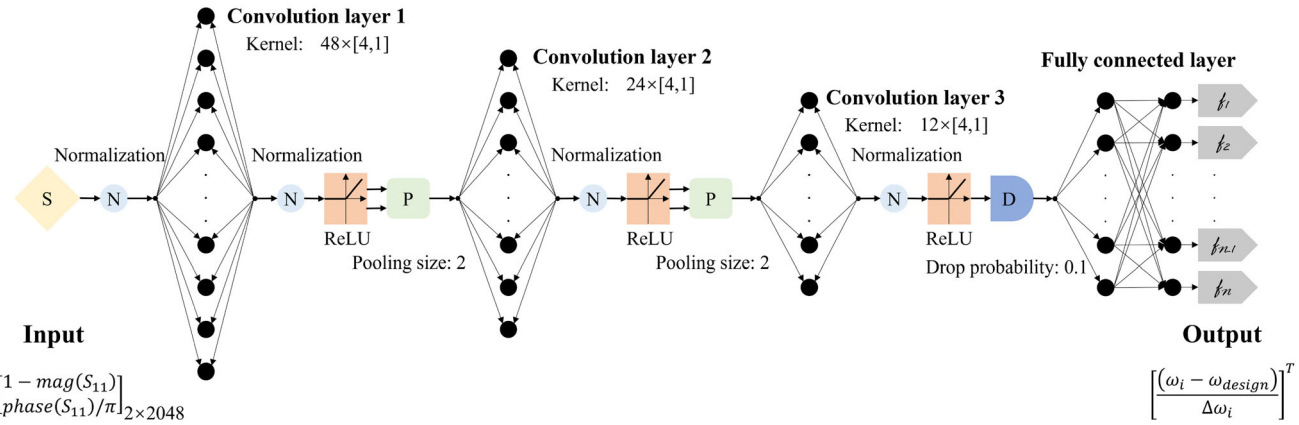


Fig. 1 (Color online) Construction of the convolutional neural network

each other, whereas values less than 0 are set to 0. This avoids the gradient explosion and gradient disappearance problems of network training and reduces the overall computational cost of neural networks. The sixth layer is a pooling layer with a pooling region size of [2,1]. The pooling layer divides the layer input into pooling regions and computes the maximum value of each region. It can reduce the dimensionality of the data and represent input information with higher-level features.

A convolutional layer, batch normalization layer, ReLU layer, and pooling layer together comprise a convolutional unit. Three of these units comprise the main body of the CNN. Through iterative testing, the number of convolution kernels for each unit was respectively tuned to 48, 24, and 12. To prevent the network from overfitting, a dropout layer is connected to the last convolutional unit. This randomly sets the layer input elements to zero, with a fixed probability of 0.1. A fully connected layer multiplies the output of the previous dropout layer by a learnable weight matrix and then adds a learnable bias vector. The fully connected layer combines all of the features and reflects them to the frequency of each cavity. To avoid the appearance of a network weight with large fluctuations, the output of the network is normalized using Eq. (2), where ω_{design} and $\Delta\omega_i$ represent the design frequency and expected detuning range of each cavity, respectively.

$$CNN_{\text{output}} = \left[\frac{\omega_i - \omega_{\text{design}}}{\Delta\omega_i}\right] \tag{2}$$

2.2 Model of the equivalent circuit

Based on the coarse outputs of the CNN, more accurate detuning can be iterated using the IPM algorithm. The gradient required for the algorithm can be calculated using the equivalent circuit model. As described by Wangler [16], an accelerator can be considered equivalent to a series of coupled circuits made up of lumped resistances, capacitances, and inductances, and the beam loading effect and RF power source can be equivalent to a voltage or current source. Each circuit obeys Kirchhoff's equation, and all the equations combine to form the equivalent matrix, as shown in Fig. 2 and Eqs. (3) and (4) for electric and magnetic coupling, respectively.

$$\begin{bmatrix} e_1 & -\frac{k_1}{2} & & & \\ & \ddots & & & \\ & -\frac{k_{i-1}}{2} & e_i & -\frac{k_i}{2} & \\ & & & \ddots & \\ & & & -\frac{k_{n-1}}{2} & e_n \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ V_{\text{e}} \\ \vdots \\ 0 \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} m_1 & -\frac{k_1}{2} & & & \\ & \ddots & & & \\ & -\frac{k_{i-1}}{2} & m_i & -\frac{k_i}{2} & \\ & & & \ddots & \\ & & & -\frac{k_{n-1}}{2} & m_n \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ V_{\text{m}} \\ \vdots \\ 0 \end{bmatrix} \tag{4}$$

In Eqs. (3) and (4), $e_i = 1 + j\frac{\omega(1+\beta)}{\omega_i Q_i} - \frac{\omega^2}{\omega_i^2}$, $m_i = 1 - j\frac{\omega_i(1+\beta)}{\omega Q_i} - \frac{\omega_i^2}{\omega^2}$, $V_{\text{e}} = 2j\omega\sqrt{C_i}\sqrt{P_0\beta R_i}$, and $V_{\text{m}} = \frac{2\sqrt{P_0\beta R_i}}{j\omega\sqrt{L_i}}$; Q_i is the quality factor of each cavity, and k_i is the coupling between two adjacent cavities. R_i, C_i, and L_i are the lumped parameters; X_i is related to the amplitude of the cavity field, $i_{\text{e},i}$ represents the beam current, and ω, P_0, and β represent the RF frequency, power, and coupling degree correlated with the feeding coupler (denoted as c).

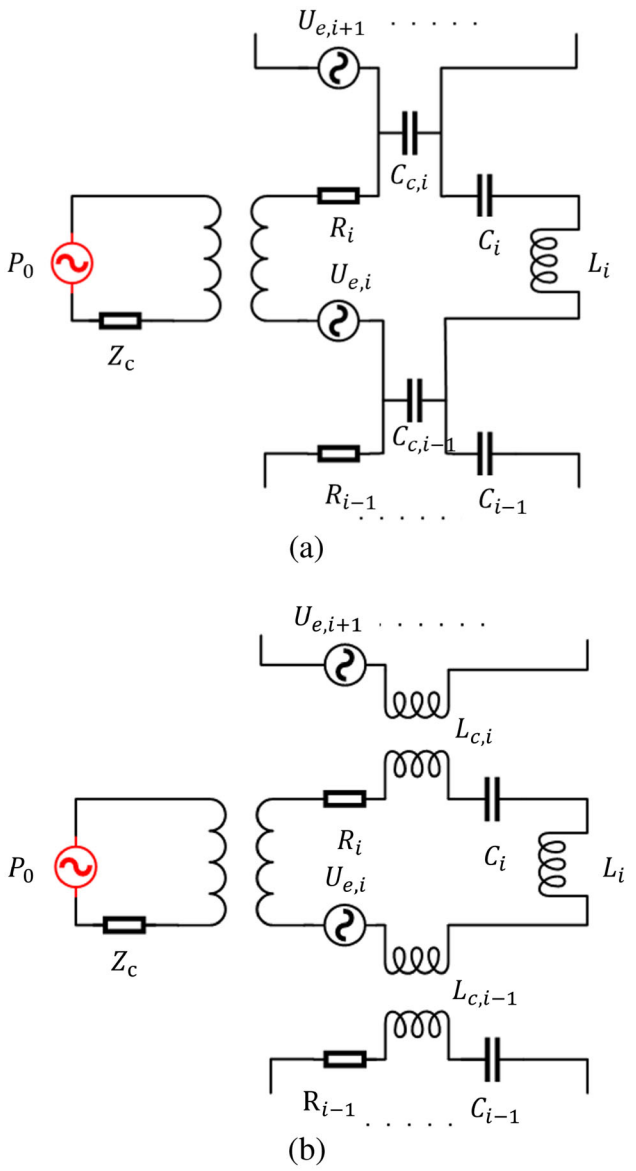


Fig. 2 Equivalent circuits for accelerators in different couplings: **a** Electric coupling; **b** magnetic coupling

Focusing on the circuit of the accelerator coupler, the coupler is equal to a voltage transformer with a ratio of n, where $n = \sqrt{\frac{\beta R_c}{Z_c}}$ and Z_c is the normalized impedance of the coupler waveguide. Assuming that the equivalent voltage of the RF power source is U_c, and the current flowing along the waveguide is I_c, one can derive the relationship between the current flowing along the waveguide and the field amplitude in the coupled cavity as $I_c = nX_c\sqrt{C_c}$. Then the normalized impedance of the accelerator coupler can be derived as $Z = \frac{U_c - I_c Z_c}{Z_c I_c} = -1 + j\frac{\omega Q_c}{\omega_c \beta X_c}$. Because the scattering parameter S_{11} satisfies $S_{11} = \frac{Z-1}{Z+1}$, the value of S_{11} measured from the accelerator coupler and its derivative with respect to the frequency of each cavity ω_i can finally be derived as

$$S_{11,\mathrm{ele}} = 1 - j\frac{2\omega\beta}{\omega_c Q_c}X_c, \quad S_{11,\mathrm{mag}} = 1 + j\frac{2\omega_c\beta}{\omega Q_c}X_c, \tag{5}$$

$$\begin{aligned}\frac{\mathrm{d}S_{11,\mathrm{ele}}}{\mathrm{d}\omega_i} &= -j\frac{2\omega\beta}{Q_c}\left(\frac{\mathrm{d}\frac{1}{\omega_c}}{\mathrm{d}\omega_i}X_c + \frac{1}{\omega_c}\frac{\mathrm{d}X_c}{\mathrm{d}\omega_i}\right),\\ \frac{\mathrm{d}S_{11,\mathrm{mag}}}{\mathrm{d}\omega_i} &= j\frac{2\beta}{\omega Q_c}\left(\frac{\mathrm{d}\omega_c}{\mathrm{d}\omega_i}X_c + \omega_c\frac{\mathrm{d}X_c}{d\omega_i}\right),\end{aligned} \tag{6}$$

where the subscripts "ele" and "mag" represent the electric and magnetic coupling structures, respectively. Abbreviate Eqs. (3) and (4) as $\mathbf{MX} = \mathbf{b}$, the expression $\frac{\mathrm{d}X_c}{\mathrm{d}\omega_i}$ is further derived as Eq. (7), where n_{ji} is the element in row j and column i of matrix $\mathbf{M}^{-1}$ and $\mathrm{d}m_{ii}$ is the diagonal element of matrix $\mathbf{M}$.

$$\frac{\mathrm{d}X_c}{\mathrm{d}\omega_i} = -n_{ji}\frac{\mathrm{d}m_{ii}}{\mathrm{d}\omega_i}X_c \tag{7}$$

Equations (5)–(7) provide the derivatives required for the IPM. For our diagnostic problem, that is, $\mathrm{argmin}(||S_{11,c}(\omega_i) - S_{11,m}||^2)$, such that $\omega_i \le \omega_{i,\mathrm{upper}}$, $\omega_i \ge \omega_{i,\mathrm{lower}}$, $i = 1, 2, ..., N$, the logarithmic penalty function $F(\omega_i)$ and its derivative with respect to the frequency of each cavity ω_i used for the IPM can be written as

$$\begin{aligned}F(\omega_i) = {} & ||S_{11,c}(\omega_i) - S_{11,m}||^2\\ & - \mu^{(k)}\sum_{i=1}^{N}(\ln(\omega_i - \omega_{i,\mathrm{lower}}))\\ & - \mu^{(k)}\sum_{i=1}^{N}(\ln(\omega_{i,\mathrm{upper}} - \omega_i)),\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\mathrm{d}F(\omega_i)}{\mathrm{d}\omega_i} = {} & 2(S_{11,c} - S_{11,m})\frac{\mathrm{d}S_{11,c}}{\mathrm{d}\omega_i}\\ & - \mu^{(k)}\sum_{i=1}^{N}\left(\frac{1}{\omega_{i,\mathrm{lower}} - \omega_i}\right)\\ & - \mu^{(k)}\sum_{i=1}^{N}\left(\frac{1}{\omega_{i,\mathrm{upper}} - \omega_i}\right),\end{aligned} \tag{9}$$

where k is the iteration number, $\mu^{(k)}$ is the penalty factor that satisfies $\mu^{(0)} > \mu^{(1)} > .....\mu^{(k)} > 0$, and $\lim_{k\to+\infty}\mu^{(k)} = 0$. The steps of the IPM are:

(1) Choose $\mu^{(0)} = 1$ as the starting penalty factor;
(2) Choose a starting point within the frequency range as described above;
(3) Optimize Eq. (8) with the derivatives calculated via Eqs. (5)–(9) until the maximal frequency error of the iteration is less than 1×10^{-3}, or the number of

iterations reaches 100. Otherwise, repeat from step(2) and multiply $\mu^{(0)} = 1$ by 0.1.

Combining the CNN with the IPM, the diagnostic steps can be summarized as follows:

(1) *Data Preparation* based on the structure designs, randomly generate groups of ω_i in a constraint zone, then calculate the associated S_{11} for each group using Eq. (5);
(2) *Network Training* divide the prepared data into a training set and validation set, and normalize the input and output arrays using Eqs. (1) and (2) to train the CNN illustrated in Fig. 1;
(3) *Coarse Estimation* for a linac to be diagnosed, measure and convert its S_{11} to the input data form, and transmit it to the trained network to estimate the coarse detuning;
(4) *Fine Calculation* further optimize the residual error $||S_{11,c} - S_{11,m}||^2$ using the IPM algorithm with the gradients determined by Eq. (6), to precisely diagnose the cavity state.

3 Results and discussion

Numerical studies were performed using three typical acceleration structures, including a 13-cell S-band standing-wave linac (SS13), a 12-cell X-band traveling-wave linac (XT12), and a 3-cell X-band RF gun (XG3). SS13 is a double-period axial coupling linac with an output beam energy of 6 MeV, XT12 is a short prototype of a constant impedance structure with 72 similar cavities used for high-gradient studies [17], and XG3 is a field-emission gun whose first cavity operates in TM_{02} mode. Although the accelerator structures selected for the numerical experiments include only S-band and X-band structures, the experimental findings can be generalized to any frequency band, because of the frequency normalization process of the CNN output layer. The inputs to both the CNN and IPM are dimensionless data, and the diagnostic results of our algorithm are related to the relative sizes of the constrained zone and bandwidth of the input frequencies. Therefore, the results can be scaled to an arbitrary frequency band. The simulation results obtained using HFSS [18] for the vacuum part and the design S_{11} of each structure are plotted in Fig. 3. The characteristics of both separated and heavily overlapped resonance peaks were considered in the experiments. 2^{13} sets of training data and 2^{10} sets of validation data were prepared for XT12 and XG3, and 2^{15} sets of training data were prepared for SS13 for better generalization performance. The network was trained using ADAM [19] with a mini-batch size of 2048 and a constant learning rate of 10^{-4}. A comparison between the S_{11} of a random validation set and the recalculated S_{11} from the diagnosis result is plotted in Fig. 4 for all three structures, and all are perfectly matched on the Smith charts. The training process for XG3 was the fastest to converge, whereas the training process for SS13 had the slowest convergence rate and largest diagnostic error. The root mean square (RMS) diagnostic error for the validation datasets for XG3, XT12, and SS13 are 160 Hz, 300 kHz, and 500 kHz, respectively. These results are related to the complexity of the accelerator structures. XG3 has the fewest cavities, whereas SS13 has the most, and the shunt impedance and quality factor vary greatly among the cavities of SS13. Further increasing the depth of the network may improve the diagnostic results for SS13; however, owing to limited computational resources, this has not yet been investigated further.

Because SS13 has the highest complexity, a comparison experiment was performed with SS13 to examine the performance for three different scenarios: diagnosis using the IPM algorithm only, using the CNN only, and using both as in Sect. 2.2. We defined the diagnostic error as the subtraction of the diagnostic cavity frequency from the corresponding cavity frequency of the validation dataset. Figure 5 shows the RMS diagnostic error histograms of each cavity obtained using the three different methods. The IPM produced the highest diagnostic error, with a maximum diagnostic error of IPM of 10 MHz. The error of the CNN has a Gaussian-like distribution, with an average value of 0 MHz and a standard deviation of approximately 1.2 MHz. The combination of the CNN and IPM achieved the best diagnostic performance. More than 95% of the combined results were accurate to within ± 500 KHz, which satisfies most engineering requirements. This comparison proves that the coarse estimate of the cavity detuning computed by the CNN successfully helps the IPM algorithm overcome local optima. The detuning estimation of the neural network in the first step and the optimization of IPM in the latter step are complementary, and neither step can accomplish the diagnosis task on its own.

Figure 6a shows a statistical histogram of the root mean square error of the SS13 diagnosis. Comparing the diagnostic accuracy of the neural network for the different cavities in Fig. 6a, it can be seen that cavity No.7, which is directly connected to the coupler, has the smallest RMS diagnostic error, whereas cavity No.1, which is the farthest away from the coupler, has the largest RMS error. The RMS error increases with the distance between the corresponding cavity and the coupler. This can be explained by coupled S-parameter calculation (CSC) theory [20]: an accelerator can be considered equivalent to a topological

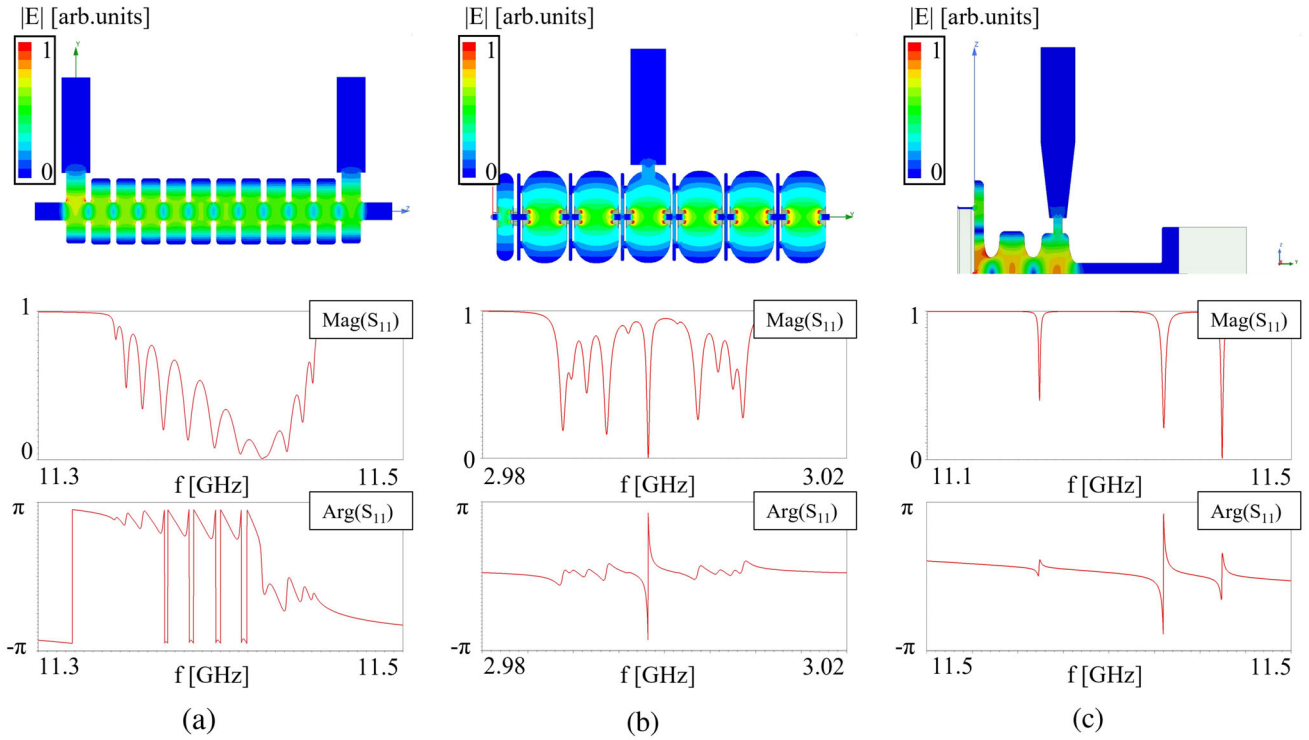


Fig. 3 (Color online) Vacuum parts, magnitude, and phase of S_{11} of the test structures: **a** XT12; **b** SS13; **c** XG3

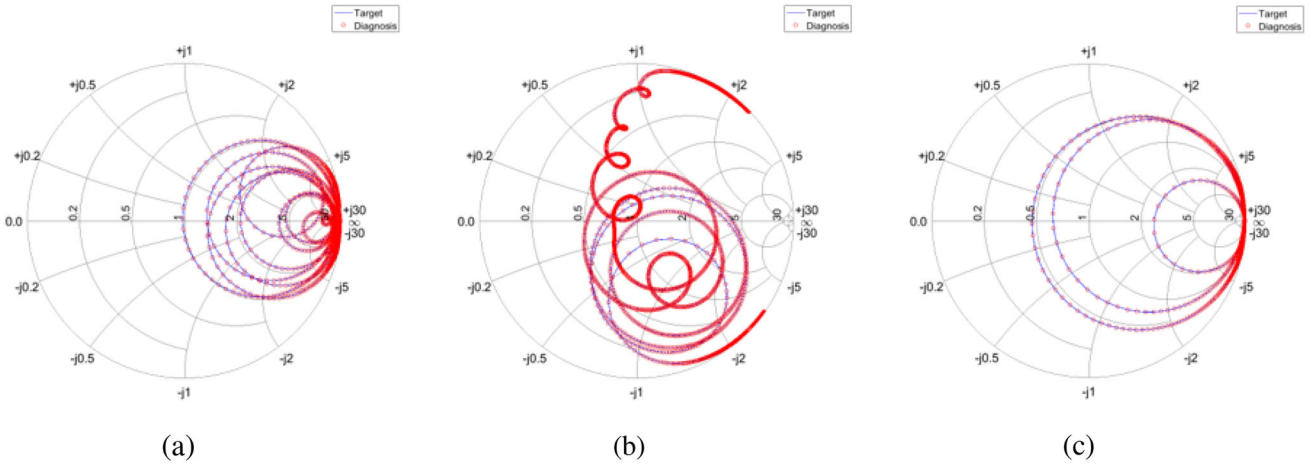


Fig. 4 (Color online) Smith chart of random validation samples of the three structures: **a** XT12; **b** SS13; **c** XG3

network constructed from a series of dual-port or triple-port units. Considering a dual-port unit as shown in Fig. 7, the sorted scattering matrix of the unit can be derived using Eq. (10), where a_i and b_i are the incident and reflected waves of each port, respectively, $S_{11,N-1}$ denotes the total scattering measured from the iris between the $(N-1)^{\text{th}}$ cavity and the N^{th} cavity, and S_{11}, S_{12}, S_{21}, and S_{22} are the scattering parameters of the isolated N^{th} cavity. According to CSC theory, the transformation matrices P and F from the intrinsic wave vector to the canonical wave vector can be derived using Eq. (11), and $S_{11,N}$ of the whole accelerator can then be written as an iteration formula as shown in Eq. (12). It can be seen from Eq. (12) that the $S_{11,N}$ of the whole accelerator is most comparable to the S_{11} of the individual coupler. The $S_{11,N}$ of the cavity farthest from the coupler is multiplied by an iteration factor of less than 1 in each topological layer. The effect of the $S_{11,N}$ of the farthest cavity on the $S_{11,N}$ of the whole accelerator becomes negligible. Therefore, the neural network will strengthen the feature extraction of the coupler cavity and surrounding cavities, while the features of the cavities far from the

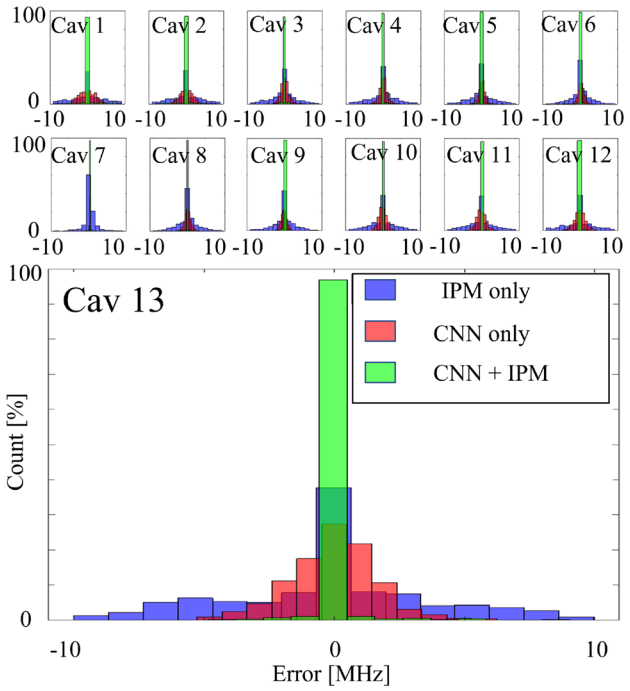


Fig. 5 (Color online) Diagnostic error histograms for each cavity of SS13. The blue squares are results from the IPM only, the red squares are results from the CNN only, and the green squares are results from the combination of the CNN and IPM

coupler are encrypted layer by layer and become more difficult to be learned by the network. Thus, for structures with additional couplers, such as traveling-wave accelerators, further addition of the scattering parameters from different couplers may improve the network performance.

$$\begin{aligned}\begin{bmatrix} a_{1,N-1} \\ a_1 \\ a_2 \end{bmatrix} &= \begin{bmatrix} S_{11,N-1} & & \\ & S_{11} & S_{12} \\ & S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_{1,N-1} \\ a_2 \\ a_1 \end{bmatrix} \\ &= \mathbf{S} \begin{bmatrix} a_{1,N-1} \\ a_2 \\ a_1 \end{bmatrix} \end{aligned} \tag{10}$$

$$\begin{bmatrix} a_{1,N-1} \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 0 & 1 \\ & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{1,N-1} \\ a_2 \\ a_1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} a_{1,N-1} \\ a_2 \\ a_1 \end{bmatrix}$$

$$\begin{bmatrix} a_{1,N-1} \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 & & 1 \\ & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_{1,N-1} \\ b_1 \\ b_2 \end{bmatrix} = \mathbf{F} \begin{bmatrix} b_{1,N-1} \\ b_1 \\ b_2 \end{bmatrix}$$

$$\mathbf{G} = \mathbf{P}^{-1}\mathbf{F}\mathbf{S}\mathbf{P} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \tag{11}$$

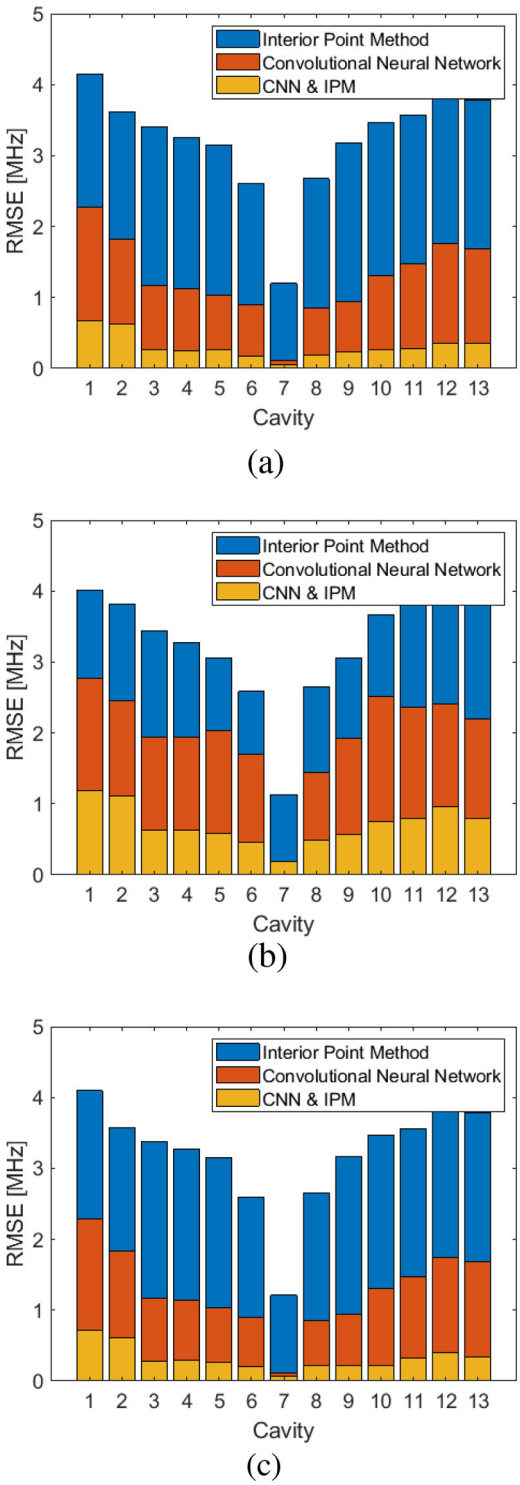


Fig. 6 (Color online) Diagnostic RMS error for each cavity with different validation sets: **a** with frequency detuning only; **b** with incorrect values of β, Q_i, k_i; **c** with data sampling noise

$$\begin{aligned} S_{11,N} &= G_{2,1}(1 - G_{1,1})^{-1} G_{1,2} + G_{2,2} \\ &= S_{11} + \frac{(S_{11,N-1} - 1) S_{12} S_{21}}{S_{22} + S_{11,N-1} - S_{11,N-1} S_{22}} \end{aligned} \tag{12}$$

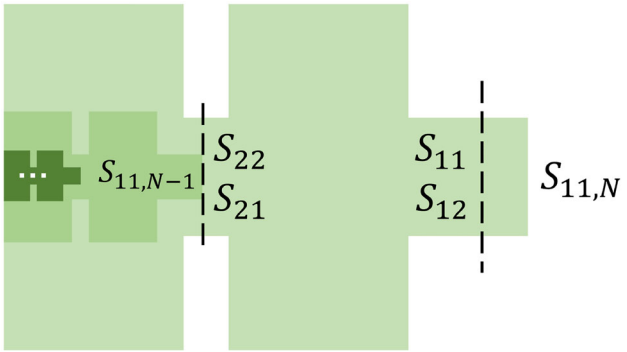


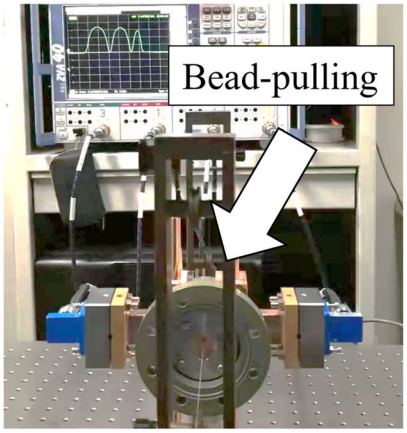
Fig. 7 (Color online) Topology of the accelerator cavity chain

To study the influence of sampling noise and errors of β, Q_i, and k_i, additional numerical experiments were conducted on SS13. Random $\Delta\beta$ generated within $[-1\%, 1\%]$, random ΔQ_i generated within $[-1.5, 1.5\%]$, and random Δk_i generated within $[-5, 5\%]$ were added to the validation data sets. Figure 6b shows the diagnostic RMS error for each cavity. When incorrect values of β, Q_i, and k_i were introduced, the RMS error of the network estimation increased. This is because the network was trained with only frequency detuning data, and the network forcibly attributed the contributions of β, Q_i, k_i to cavity frequency detuning. However, the IPM can still converge to a better diagnosis value based on the network output. In addition, a sampling noise of Gaussian distribution with a signal-to-noise ratio of 60 dB was added to the validation datasets, while the network was still trained with noiseless datasets. Figure 6c shows the RMS diagnostic errors. As shown by the red bins in Fig. 6c, the output error has a similar distribution between the non-noisy dataset and the noise-added dataset. This implies that the network is partially resistant to sampling noise, probably because of the pooling layers. The pooling layers serve to downsample the data to reduce the computational cost of the neural network. During this process, the data can be considered to have passed through a low-pass filter, which has the effect of filtering out noise.

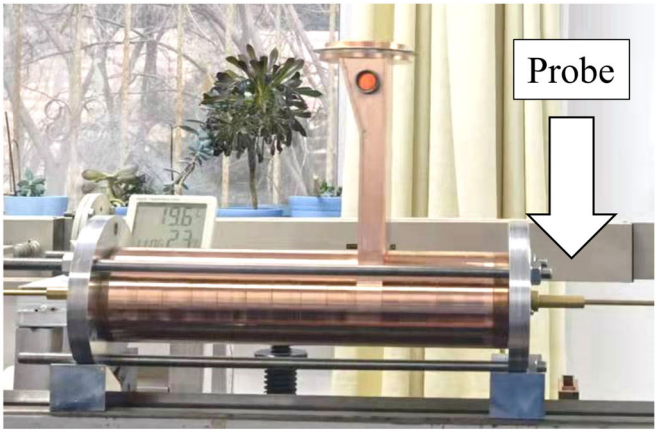
A well-trained network for XG3 was used to guide the tuning procedure of a real gun, as shown in Fig. 8a. A testing cathode with a hole in its center was installed on the gun for bead-pulling. The field distributions of the three different working modes were measured to apply the diagnostic method based on field distributions. Diagnosis results from our method were compared with those based on the field distribution. As shown in Fig. 8c, the frequency diagnosis differences between the two methods for the first and third cavities are 190 KHz and 230 KHz, respectively, whereas the difference for the second cavity is 3.2 MHz. This is because the field distributions in the second cavity were almost zero in $\pi/2$ mode, resulting in a singular matrix in the bead-pulling method. A slight difference or jittering of the pulling string may introduce significant measurement errors into the field distribution of the second cavity. Therefore, it is difficult for a diagnostic method based on field distributions to compute the frequency of the second cavity accurately. Under the guidance of our new method, the π mode frequency of the gun was tuned to 11.424 GHz, and the on-axis electric field distribution was consistent with the design value, as shown in Fig. 8e. The new diagnostic method completely satisfies the tuning requirements of the gun cavity. Another network was trained to help tune a standing-wave linac (SS13-2). SS13-2 is similar to SS13, but differs in the coupler position. The coupler position of SS13-2 was at the 9th cavity, whereas that of SS13 was at the 7th. The comparison shown in Fig. 8d was made between the diagnosis results of our method and those of the probe insertion method. The difference between the two was less than 500 KHz. Figure 8d shows that the results of our method fluctuate around those obtained by the probe insertion method. This can be explained in the same way as the difference in the numerical experiments with deliberately incorrect values of β, Q_i, and k_i. The actual values of these parameters for the tested SS13-2 were slightly different from the design values, adding a frequency shift to the diagnostic results. Including the parameter values in the IPM step may eliminate this frequency shift. Figure 8f shows a comparison of $S_{11,\mathrm{m}}$ and $S_{11,\mathrm{c}}$ for SS13-2. Owing to these factors, there is also a small deviation between the two lines. However, for our engineering needs, this error was negligible. In summary, these two results prove that the combined CNN and IPM diagnostic method is in good agreement with both conventional methods and has a major advantage in that the detuning data can be obtained in almost real time after training. After training with prepared data sets, one can obtain the accelerator cavity detuning information immediately, while other methods require time for field measurements or probe adjustments. As the tuning process needs to be repeated several times, the conventional methods may take several hours of processing time in total, but the combined CNN and IPM method can reduce the processing time to a few minutes.

4 Conclusion

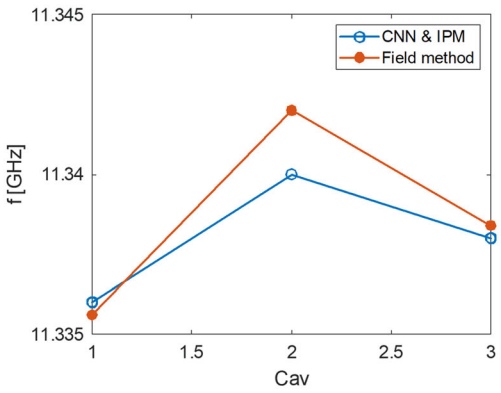
In conclusion, we developed a non-invasive diagnostic method for cavity detuning. This approach first trains a convolutional neural network to estimate the frequency detuning of each cavity with the input of the measured S_{11} and then uses this estimation value as the starting point for the IPM algorithm to further optimize the divergence



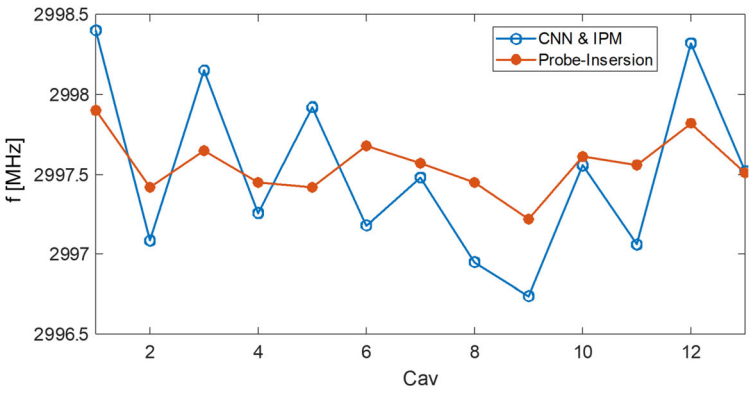
(a)



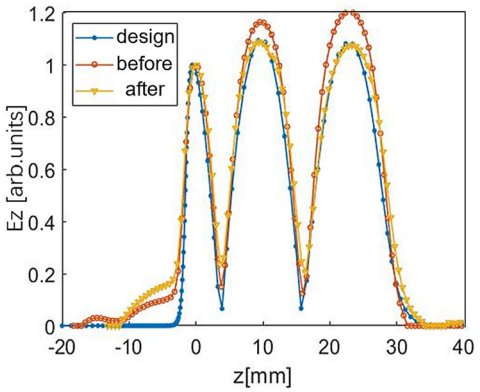
(b)



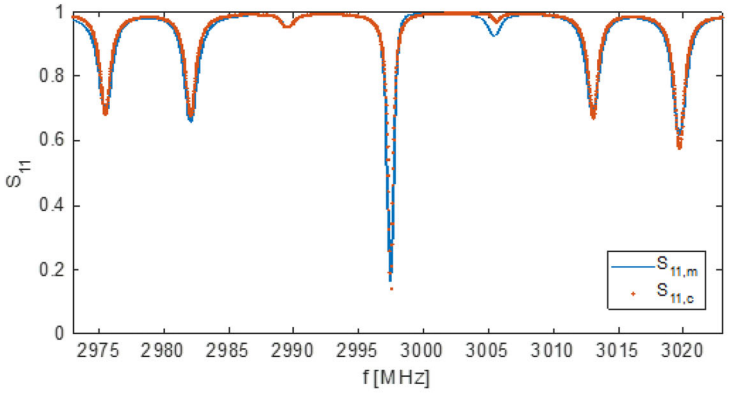
(c)



(d)



(e)



(f)

Fig. 8 (Color online) Testing with real structures: **a** XG3 gun; **b** SS13-2 linac; **c** benchmark comparison between the CNN & IPM method and the field method; **d** benchmark comparison between the CNN & IPM method and the probe insertion method; **e** on-axis electric field distribution of XG13; **f** comparison between the $S_{11,c}$ and $S_{11,m}$ of SS13-2

between the calculated S_{11} using equivalent circuits and the measured S_{11}. The convolutional neural network has a total of 15 layers. The 3rd, 7th, and 11th layers are convolutional layers with 48, 24, and 12 kernels, respectively. The network was trained using simulation datasets generated from the equivalent circuits. Numeric experiments were

successfully completed on three different acceleration structures, including a 13-cell S-band standing-wave linac, a 12-cell X-band traveling-wave linac, and a 3-cell X-band RF gun. Owing to the topological nature of the structure, the diagnostic accuracy of this method decreases as the distance from the cavity to the coupler increases. This method is robust to sampling noise owing to the use of pooling layers. The well-trained network also aided in tuning real structures. The diagnostic results of this method were in good agreement with those of conventional methods.

This approach provides a fresh perspective on the diagnosis of high-frequency bands, long cavity chains, and encapsulated accelerators. After hours of pre-training, detuning information can be obtained in situ simply by measuring the S_{11} parameters. We anticipate that this method will significantly reduce the burden of the tuning process and provide a new approach for monitoring the status of encapsulated linacs. In future work, we will continue to tune the structure of the network and attempt to include other accelerator parameters in the diagnostic algorithm to enable the diagnosis of more complex acceleration structures.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Liu-Yuan Zhou, Jia-Ru Shi. The first draft of the manuscript was written by Liu-Yuan Zhou, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. X. Pu, H. Hou, Y. Wang et al., Frequency sensitivity of the passive third harmonic superconducting cavity for SSRF. Nucl. Sci. Tech. **31**, 176 (2020). https://doi.org/10.1007/s41365-020-0732-x
2. H. Shi, H. Ouyang, S. Wang et al., RF tuning and beam commissioning of CW RFQ for china-ads injector-*i*. Nucl. Sci. Tech. **29**, 142 (2018). https://doi.org/10.1007/s41365-018-0478-x
3. R. Neal, J.P. Blewett, The Stanford two-mile accelerator. Phys. Today **23**, 76 (1968). https://doi.org/10.1063/1.3022031
4. T. Khabiboulline, V. Puntus, R.M. Dohlus et al., A new tuning method for traveling wave structures. **3**, 1666–1668 (1995). https://doi.org/10.1109/PAC.1995.505321
5. J. Shi, A. Grudiev, A. Olyunin et al., Tuning of CLIC accelerating structure prototypes at CERN. LINAC2010, 03 (2022)
6. Y. Yang, J. Yang, X. Wang et al., A quantitative calculation method of RF parameters for traveling wave accelerating structures. Nucl. Instrum. Methods Phys. Res., Sect. A **989**, 16492 (2020). https://doi.org/10.1016/j.nima.2020.164923
7. E. Häbel and J. Tückmantel. Tuning of a superconducting accelerating cavity under operating conditions: Part 1 : theory. 01 (1981)
8. Y. Ni, D. Tong, Y. Lin, Genetic algorithm diagnosis of individual cell frequencies in a coupled cavity chain. Nucl. Instrum. Meth. Phys. A **462**, 356–363 (2001). https://doi.org/10.1016/S0168-9002(00)01327-9
9. A. Ivanov, I. Agapov, Physics-based deep neural networks for beam dynamics in charged particle accelerators. Phys. Rev. Accel. Beams **23**, 074601 (2020). https://doi.org/10.1103/PhysRevAccelBeams.23.074601
10. Y. Zou, Q. Xing, B. Wang et al., Application of the asynchronous advantage actor-critic machine learning algorithm to real-time accelerator tuning. Nucl. Sci. Tech. **30**, 158 (2019). https://doi.org/10.1007/s41365-019-0668-1
11. X. Ren, A. Edelen, A. Lutman et al., Temporal power reconstruction for an x-ray free-electron laser using convolutional neural networks. Phys. Rev. Accel. Beams **23**, 040701 (2020). https://doi.org/10.1103/PhysRevAccelBeams.23.040701
12. V. Kain, S. Hirlander, B. Goddard et al., Sample-efficient reinforcement learning for CERN accelerator control. Phys. Rev. Accel. Beams **23**, 124801 (2020). https://doi.org/10.1103/PhysRevAccelBeams.23.124801
13. J.E. Rayas-Sanchez, Artificial neural networks and space mapping for EM-based modeling and design of microwave circuits, in *Surrogate-Based Modeling and Optimization*. ed. by S. Koziel, L. Leifsson (Springer, New York, NY, 2013), pp. 147–169. https://doi.org/10.1007/978-1-4614-7551-4_7
14. J. Goncalves, R. Storer, J. Gondzio, A family of linear programming algorithms based on an algorithm by von Neumann. Optimiz. Methods Softw. **24**, 461–478 (2009). https://doi.org/10.1080/10556780902797236
15. X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statisitics (AISTATS) 2011 **15**, 315–323 (2011)
16. T. Wangler, RF linear accelerators. Second Edition (2008). https://doi.org/10.1002/9783527623426
17. J. Liu, J. Shi, H. Zha et al., Analytic RF design of a linear accelerator with a sled-I type RF pulse compressor. Nucl. Sci. Tech. **31**, 107 (2020). https://doi.org/10.1007/s41365-020-00815-5
18. Ansys HFSS best-in-class 3d high frequency electromagnetic simulation software
19. D. Kingma, J. Ba, Adam: A method for stochastic optimization. International Conference on Learning Representations, **12**, (2014)
20. T. Flisgen, E. Gjonaj, G. Walter et al., Generalization of coupled s -parameter calculation to compute beam impedances in particle accelerators. Phys. Rev. Accel. Beams **23**, 034601 (2020). https://doi.org/10.1103/PhysRevAccelBeams.23.034601