Title	Machine learning-based prediction of the electron energy distribution function and electron density of argon plasma from the optical emission spectra
Author(s)	Arellano, Fatima Jenina; Kusaba, Minoru; Wu, Stephen et al.
Citation	Journal of Vacuum Science and Technology A. 2024, 42(5), p. 053001
Version Type	VoR
URL	https://hdl.handle.net/11094/97738
rights	This article may be downloaded for personal use only. Any other use requires prior permission of the author and AIP Publishing. This article appeared in Arellano F.J., Kusaba M., Wu S., et al. Machine learning-based prediction of the electron energy distribution function and electron density of argon plasma from the optical emission spectra. Journal of Vacuum Science and Technology A 42, 053001 (2024) and may be found at https://doi.org/10.1116/6.0003731.
Note	

RESEARCH ARTICLE | JULY 11 2024

Machine learning-based prediction of the electron energy distribution function and electron density of argon plasma from the optical emission spectra ⊖⊘

Special Collection: Artificial Intelligence and Machine Learning for Materials Discovery, Synthesis and Characterization

Fatima Jenina Arellano ⓘ ; Minoru Kusaba ⓘ ; Stephen Wu ⓘ ; Ryo Yoshida ⓘ ; Zoltán Donkó ⓘ ; Peter Hartmann ⓘ ; Tsanko V. Tsankov ⓘ ; Satoshi Hamaguchi ✉ ⓘ

Check for updates

CHORUS

View Online

Export Citation

19 August 2024 04:59:59

Machine learning-based prediction of the electron energy distribution function and electron density of argon plasma from the optical emission spectra EP

Fatima Jenina Arellano,[1] [ID] Minoru Kusaba,[2] [ID] Stephen Wu,[2,3] [ID] Ryo Yoshida,[2,3] [ID] Zoltán Donkó,[1,4] [ID]
Peter Hartmann,[4] [ID] Tsanko V. Tsankov,[5,6] [ID] and Satoshi Hamaguchi[1,a)] [ID]

AFFILIATIONS

[1]Graduate School of Engineering, Osaka University, Osaka 565-0871, Japan
[2]The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa 190-8562, Japan
[3]Department of Statistical Science, The Graduate University for Advanced Studies, Tachikawa 190-8562, Japan
[4]Institute for Solid State Physics and Optics, HUN-REN Wigner Research Centre for Physics, Budapest 1121, Hungary
[5]Faculty of Physics and Astronomy, Experimental Physics V, Ruhr University Bochum, Bochum, 44801, Germany
[6]LPP, CNRS, Sorbonne Université, École Polytechnique, Institut Polytechnique de Paris, Palaiseau 91128, France

Note: This paper is part of the Special Topic Collection on Artificial Intelligence and Machine Learning for Materials Discovery, Synthesis and Characterization.
[a)]Author to whom correspondence should be addressed: hamaguch@ppl.eng.osaka-u.ac.jp

ABSTRACT

Optical emission spectroscopy (OES) is a highly valuable tool for plasma characterization due to its nonintrusive and versatile nature. The intensities of the emission lines contain information about the parameters of the underlying plasma–electron density n_e and temperature or, more generally, the electron energy distribution function (EEDF). This study aims to obtain the EEDF and n_e from the OES data of argon plasma with machine learning (ML) techniques. Two different models, i.e., the Kernel Regression for Functional Data (KRFD) and an artificial neural network (ANN), are used to predict the normalized EEDF and Random Forest (RF) regression is used to predict n_e. The ML models are trained with computed plasma data obtained from Particle-in-Cell/Monte Carlo Collision simulations coupled with a collisional–radiative model. All three ML models developed in this study are found to predict with high accuracy what they are trained to predict when the simulated test OES data are used as the input data. When the experimentally measured OES data are used as the input data, the ANN-based model predicts the normalized EEDF with reasonable accuracy under the discharge conditions where the simulation data are known to agree well with the corresponding experimental data. However, the capabilities of the KRFD and RF models to predict the EEDF and n_e from experimental OES data are found to be rather limited, reflecting the need for further improvement of the robustness of these models.

I. INTRODUCTION

Low-temperature plasma has emerged as a versatile technology with diverse applications across various sectors, including semiconductor processing, electric propulsion, ozone generation, and plasma agriculture and medicine.[1–3] One crucial factor in achieving significant progress in these domains lies in the ability to precisely characterize and control the plasma required for these applications. For this reason, over several decades, significant efforts have been placed into developing plasma diagnostics.[4,5]

In certain applications, such as in semiconductor manufacturing and in plasma propulsion systems, it is desirable to be able to determine the plasma characteristics without disturbing the plasma. This necessitates using noninvasive diagnostics such as spectroscopy

methods, laser-based methods,[4,6] and wall-mounted measurement systems.[5,7] Optical emission spectroscopy (OES) is a widely used noninvasive plasma diagnostic valued for its versatility and simplicity as compared to other diagnostic tools. It is extensively employed in characterizing plasma composition and tracking dynamic changes in the plasma.[8,9] Moreover, OES can be employed to determine plasma parameters, such as the electron density n_e and the electron energy distribution function (EEDF) or the electron temperature T_e, as the electron kinetics are reflected in the excitation dynamics. There exist a number of methods to determine n_e and EEDF from the optical emission spectra. One such method relies on determining the excitation temperature T_{exc} from the slope of the Boltzmann plot of ln $(I\lambda/g_2 A_{12})$ vs the energy of the higher excited level, wherein I, λ, g_2, and A_{12} are the intensity, wavelength, statistical weight, and Einstein A-coefficient of the corresponding transition. The electron temperature T_e can be set equal to T_{exc} at local thermal equilibrium (LTE) or can alternatively be determined using the Saha equation and the line-continuum ratio. The electron density n_e is then determined using the measurement of the Stark broadening of the lines.[10–12] However, such a method assumes that the plasma is in local thermal equilibrium, which typically exists only at high electron densities and high pressures.[10]

Another approach involves assuming a corona regime and relating T_e to the ratio of the emission line intensities from two excited levels. In this method, one line typically originates from an ionic level, while the other comes from a neutral level of the same atomic species.[10,13,14] However, the assumption of a corona model is usually valid only at low pressures.[15,16] If trace gases are added to the plasma, actinometry and trace rare gas OES can also be used to determine the density of ground-state species and T_e.[10,17]

Another method that is commonly used to relate spectral line intensities to plasma parameters employs so-called collisional-radiative models (CRMs). A CRM takes into account the processes that populate and depopulate each excited level to solve a rate balance equation for these levels. In this way, plasma parameters such as n_e and the EEDF can be related to the rate of spontaneous emission reflected in the intensities of spectral lines. One way a CRM is often utilized is by varying n_e and T_e until the resulting intensities agree closely with the measured OES intensities.[15,18–25] However, such calculations usually need an assumption on the distribution of the electron energies. The usual assumptions of a single-temperature Maxwellian or two-temperature distributions are often not accurate enough to represent the actual conditions.[16] A more straightforward way to use the model is to employ n_e and the EEDF as input parameters to the CRM in order to yield the corresponding spectral intensities. This way, "inverting" the CRM would lead to the determination of the plasma parameters from the OES spectra. However, solving for the inverse of the CRM is often challenging due to the complexity of determining the realistic shape of the EEDF represented with a limited number of parameters.

Recently, there have also been a growing number of studies employing machine learning (ML) to determine the plasma parameters from the OES line intensities.[26–29] Shojaei and Mangolini used the OES and Langmuir probe measurements to predict the electron energy probability function (EEPF) from normalized emission spectral intensities.[28] van der Gaag et al. used a genetic

algorithm to predict regions of the EEDF from the emission from electron-atom bremsstrahlung.[29] Park et al. also predicted T_e and n_e from the OES intensities.[27]

The final goal of this study is to develop ML models to predict the (normalized) EEDF and n_e of Ar plasma from its OES data under various discharge conditions. Because the total gas pressure of the plasma system can be measured easily, it can also be used as an input parameter. The EEDF may not be represented by a combination of Maxwell distributions or other known distribution functions, so we do not assume any particular parametrized functional forms for the EEDF. We consider a plasma system with limited diagnostic accessibility and assume that the optical emission is measured at a single point in space and the absolute value of the photon flux cannot be determined. Therefore, the relative (i.e., normalized) emission line intensities accumulated over a certain time interval, rather than the absolute emission line intensities, are used as input data for the ML models. Because the observed optical emission signals are integrated along the line of sight, the nonuniformity of the plasma can also affect the observed line intensities. The assumption is then made that the relative emission line intensities are related to the electron density and EEDF averaged over the line of sight and that this information can be extracted by the ML models.

If such ML models are successfully developed and the prediction is fast enough, real-time monitoring of the plasma density and EEDF of Ar plasmas can be achieved based on the real-time information of measured OES data and gas pressure. Furthermore, if the ML models developed in this study can be extended to chemically reactive plasmas, they can be used as real-time monitoring tools for industrial plasma systems.

In this study, as the first step toward the final goal, we developed such ML models using numerical simulation data as their training data. The decision to use simulation data for training our machine learning models is driven by the need for noninvasive diagnostics and a higher dynamic range of the EEDF. Although training on experimental data might seem ideal for developing accurate predictive models, it is important to recognize that each diagnostic method comes with its own uncertainties and biases, which can be imprinted onto the model. For instance, Langmuir probes tend to be invasive and their applicability becomes questionable at elevated pressures. Thomson scattering is challenging at low densities and requires long measurement times, making the collection of sufficient data for the training of the models unrealistic. Finally, all experimental methods lack the sensitivity to cover the large dynamic range of variation of the EEDF while models such as the particle-in-cell/Monte Carlo collision (PIC/MCC) simulation readily provide this capability.

The numerical simulation data for training are those of Ar capacitively coupled plasmas obtained from the simulation codes described in Ref. 16. The prediction capability of the developed ML models is tested with the simulated OES data as their input parameters at various gas pressures. The simulation data used for our ML training are known to agree well with the corresponding experimental data when the gas pressure is about 20 Pa or lower although some discrepancy always exists between the simulated and experimentally measured OES data. The discrepancy can arise from various reasons, ranging from the inaccuracy of the simulation model to the presence of impurities and nonuniformity in the

actual plasma used for the measurement. The presence of a small amount of impurities, such as nitrogen, could already cause notable quenching of the argon excited states. In addition, the simulation models assume a uniform profile of the plasma along the line of sight; therefore, any variations in plasma uniformity add to prediction errors in the simulated data and in the ML models. However, if the developed ML models are robust enough, they should be able to predict the electron density and EEDF of the actual plasma at a pressure of up to 20 Pa from experimental OES data. The robustness of the ML models was also tested in this study.

This paper is divided into five parts. The preparation of training and test data is outlined first in Sec. II, followed by an introduction to the ML models used in Sec. III. Subsequently, the prediction results for simulated data are then presented and discussed in Sec. IV, followed by those for experimental data in Sec. V. Finally, the concluding remarks are given in Sec. VI.

II. DATA PREPARATION

This study used the intensities of 15 spectral lines as the input vector to predict the EEDF and n_e as the output data. In addition to the OES intensities, the pressure was also included as an input variable, resulting in a 16-dimensional input vector. The details of the spectral lines used in this study are provided in Table I. The functional variable EEDF was discretized on an equidistant grid having 1999 points, while n_e was given as a discrete scalar variable. First, the preparation of the input vectors from calculated and measured data is outlined in the following.

A. PIC+CRM simulation data

The simulation dataset used for ML was generated through the one-way coupling of a one-dimensional (1D) PIC/MCC model and a CRM, denoted as "PIC+CRM,"[16] of capacitively coupled plasmas (CCPs) in argon (Ar) gas. The CRM was used to compute

TABLE I. List of the selected argon lines used in the ML study. The levels of the transitions are denoted in Paschen notation and the data are taken from Ref. 30.

Wavelength (nm)	Lower level	Upper level
696.543	$1s_5$	$2p_2$
706.722	$1s_5$	$2p_3$
714.740	$1s_5$	$2p_4$
727.294	$1s_4$	$2p_2$
738.398	$1s_4$	$2p_3$
750.387	$1s_2$	$2p_1$
751.465	$1s_4$	$2p_5$
763.510	$1s_5$	$2p_6$
772.376	$1s_5$	$2p_7$
772.422	$1s_3$	$2p_2$
794.818	$1s_3$	$2p_4$
800.616	$1s_4$	$2p_6$
801.479	$1s_5$	$2p_8$
810.369	$1s_4$	$2p_7$
811.531	$1s_5$	$2p_9$
826.452	$1s_2$	$2p_2$

the OES spectral line intensities corresponding to a specific combination of EEDF and n_e, both calculated using PIC/MCC simulations. The same PIC+CRM model described in detail in Ref. 16 was employed in this work. In this previous work, only the pressure was varied between 2 and 100 Pa at a constant peak-to-peak voltage of 300 V. Here, data for other voltages between 200 and 500 V (peak-to-peak) for the same pressure values have also been generated. In this way, a total of 108 sets of OES intensities and plasma parameters were generated for the training and testing of the ML models.

The CRM used here[16] was derived from the work of Ref. 21. It takes into account the lowest 14 excited levels of argon, namely, the 1s and the 2p levels (in Paschen's notation). A CRM has several input parameters, including the collision cross sections, the parameters of the chamber geometry, a list of radiative transitions considered and a set of coefficients corresponding to the processes involved. In this work, these parameters were set in accordance with the specifications shown in Ref. 16. More specifically, the zero-dimensional CRM was computed for an argon plasma within a cylindrical chamber enclosing highly symmetrical parallel-plane electrodes. Each electrode had a diameter of 14.2 cm, and the gap distance between them was set at 4 cm. For each pair of input and output data, all other parameters are the same except for the EEDF n_e, the gas temperature T_g, and the pressure. The values used for T_g ranged from 304 to 348.4 K and were interpolated from experimental tunable diode laser absorption spectroscopy (TDLAS) measurements done in Ref. 16. The resulting intensities calculated from the CRM were given in units of photons, $m^{-3} s^{-1}$.[16,31]

The EEDF and n_e were calculated using a one-dimensional (1D) electrostatic PIC/MCC model.[32,33] The simulations were conducted for a capacitively coupled radio frequency argon plasma with the same chamber configurations described above. The EEDF and n_e values utilized in the ML study were averaged over the central 1 cm diameter region of the plasma to align with the region of measurement in the OES experiments detailed in Sec. II B. The PIC/MCC simulations also employed the same interpolated T_g values as described above. Before being employed in the CRM, the EEDF f_e calculated from PIC/MCC simulations was normalized over the electron energy ε such that[16,21]

$$\int_0^\infty f_e(\varepsilon)d\varepsilon = 1. \tag{1}$$

In the same way, the predicted EEDFs were normalized according to Eq. (1) before being used in the CRM.

Out of the 108 data sets that were generated, 80% were randomly selected for training, and the remaining samples were reserved for testing. The logarithm of the EEDFs was taken and these resulting EEDFs were used as the training data as it was observed that predictions in linear scale tend to neglect the higher energy region. Additionally, to ensure that even zero values could be transformed into the logarithmic scale, a small constant value of $10^{-10} eV^{-1}$ was added to each data point in the EEDF. For comparison with experimental intensities, the intensities calculated from the CRM were normalized to the total intensity of the selected

spectral lines such that

$$\tilde{I} = I_k / \sum I_k, \tag{2}$$

where I_k denotes the intensity of each selected argon line and $\tilde{I}$ represents the resulting normalized intensity. It should be noted that the intensities at 772.38 and 772.42 nm were not well-resolved in the measurements. Therefore, in order to compare with the experiments, the two intensities calculated from the CRM were combined into a single value corresponding to an effective wavelength of 772.4 nm.

B. Experimental data

The experimental data used in this study were taken from Ref. 16. The readers are, therefore, referred to this paper for more details. Briefly, the experimental measurements were conducted on an argon plasma employing the same chamber configurations described in the simulations outlined in Sec. II A. However, the experimental measurements were limited to a single peak-to-peak voltage setting, specifically at 300 V. Three sets of measurement data were incorporated in this study: intensities obtained from the OES, gas temperature T_g, and $1s_5$ density measured using tunable diode laser absorption spectroscopy (TDLAS).

The experimentally measured OES intensities served as a second set of test input data for the evaluation of the predictive performances of the ML models. Similar to the treatment applied to calculated intensities, the measured relative intensities were also normalized using Eq. (2). The measured T_g data were interpolated and employed as input parameters for both the PIC/MCC simulations and the CRM. On the other hand, the measured $1s_5$ density data served as reference data for additional validation of the predicted results.

III. ML MODELS

Three different ML models were used in this study—two for EEDF prediction and one for predicting n_e. Specifically, the Kernel Regression for Functional Data (KRFD) and Artificial Neural Network (ANN) models were employed to predict the EEDF, while n_e was predicted using an Random Forest (RF) regressor. The hyperparameters for each model were selected by Bayesian optimization using Optuna Python library[34] separately for each model, with the number of trials set to 300. Fivefold cross-validation was employed to choose the validation set and calculate the validation loss. Fivefold cross-validation is an ML technique to assess the model performance where the dataset (excluding the test subset) is divided randomly into five equal parts. The model is then trained and tested five times, each time using a different part as the test set and the remaining parts as the training set.

A. KRFD model

The KRFD was originally developed in Ref. 35. The method allows to obtain a function $Y(t)$ from an input vector X. The predicted output $Y(X, t)$ is modeled by

$$Y(X, t) = \sum_{i=1}^{T} \beta_i(X) k_T(t, t_i) + \mu(t) + \delta. \tag{3}$$

Here, T is the total number of data points in t. In this study, t is set to be the electron energy, $Y(t; X)$ is assigned as the EEDF at that energy, while X represents the 16-dimensional input vector comprising OES intensities and pressure. A set of T positive definite kernel functions, $k(t, t_i)$ ($i = 1, \ldots, T$), is placed at arbitrarily specified centers t_i ($i = 1, \ldots, T$) within the support of the electron energy. The regression coefficient or weight parameter $\beta_i(X)$ is modeled as a function of X, which activate or deactivate each kernel. While in Ref. 35, the coefficients were modeled with ANNs, in this study kernel regressors are used to achieve more robust and cost-effective modeling. Additionally, $\mu(t)$ is a baseline function that relies only on t and δ is the noise term. Unlike ordinary regression, where the output is typically given as a scalar variable or relatively low-dimensional vector, the output variable here is treated as a function of t. Such a task is called functional output regression. For the kernels $k(t, t_i)$, we use the Gaussian radial basis kernel function (RBF) given by

$$k(t, t_i) = l \, \exp\left(-\frac{\|t - t_i\|^2}{2\sigma^2}\right), \tag{4}$$

where l is a length scale parameter, σ is the bandwidth parameter of the basis function, and $\|t - t_i\|$ is the Euclidean distance of the electron energy t from the ith kernel center t_i. While the use of other RBF can be explored as a future work, a Gaussian RBF was chosen due to its wide usage as a kernel function in kernel regression models.

The coefficient functions $\beta_i(X)$ ($i = 1, \ldots, T$) are modeled with additional kernel regressors. Specifically, Eq. (3) is expressed as

$$Y(X, t) = \sum_{k=1}^{N} \sum_{l=1}^{T} \theta_{kl} k_G(X, X_k) k_T(t, t_l) + \sum_{m=1}^{N} c_m k_M(X, X_m) + \delta. \tag{5}$$

Here, θ_{kl} and c_m represent the regression coefficients, $k_G(X, X_k)$ and $k_M(X, X_m)$ denote the Gaussian RBF kernels on variable X, $k_T(t, t_l)$ is the RBF kernel on t, and N denotes the number of the kernel functions for X. This can be obtained by setting $\beta_i(X) = \sum_{k=1}^{N} \theta_{kl} k_G(X, X_k)$ in Eq. (3).

In this study, ridge regression with ℓ_2 regularization was performed to estimate the coefficient parameters, θ_{kl} and c_m, where the objective function O to be minimized during model training is given by

$$O(\boldsymbol{\Theta}, \boldsymbol{c}) = \|\boldsymbol{Y} - \boldsymbol{G\Theta T} - \boldsymbol{Mc1}^T\|_F^2 + \alpha_G \|\boldsymbol{G}^{1/2}\boldsymbol{\Theta T}\|_F^2 + \alpha_T \|\boldsymbol{G\Theta T}^{1/2}\|_F^2$$
$$+ \alpha_G \alpha_T \|\boldsymbol{G}^{1/2}\boldsymbol{\Theta T}^{1/2}\|_F^2 + \alpha_M \|\boldsymbol{M}^{1/2}\boldsymbol{c1}^T\|_F^2, \tag{6}$$

where $\boldsymbol{Y}$, $\boldsymbol{G}$, $\boldsymbol{\Theta}$, $\boldsymbol{T}$, and $\boldsymbol{M}$ are $N \times T$, $N \times N$, $N \times T$, $T \times T$, and $N \times N$ matrices whose entries are given by $\boldsymbol{Y}_{ij} = Y(X_i, t_j)$,

$G_{ij} = k_G(X_i, X_j)$, $\boldsymbol{\Theta}_{kl} = \theta_{kl}$, $\boldsymbol{T}_{ij} = k_T(t_i, t_j)$, $\boldsymbol{M}_{ij} = k_M(X_i, X_j)$, and $\boldsymbol{c} = (c_1, \ldots, c_N)^{\mathrm{T}}$. Here, $\boldsymbol{1}$ denotes a T-dimensional vector whose entries are all one and a square matrix raised to the power $1/2$ is another square matrix whose product with itself gives the original matrix. Additionally, $\| \ \|_F$ represents the Frobenius norm. The optimal solutions for $\boldsymbol{\Theta}$ and $\boldsymbol{c}$ were determined by solving the system of equations obtained by setting $\frac{\partial O(\boldsymbol{\Theta}, \boldsymbol{c})}{\partial \boldsymbol{\Theta}} = 0$ and $\frac{\partial O(\boldsymbol{\Theta}, \boldsymbol{c})}{\partial \boldsymbol{c}} = 0$.

Each of the 16 input variables was standardized to have a mean of 0 and a standard deviation of 1 in the training dataset. Note that this procedure does not preserve the line ratios within a given spectrum. About 20% of the total data was randomly selected for the test set, and the remainder was used as the training set. The hyperparameters were selected using Optuna through fivefold cross-validation within the training set. The parameters to be selected were σ_G, σ_T, and $\sigma_M \in [1, 100]$, which are the bandwidth parameters for the Gaussian RBF kernels $k_G(X, X_k)$, $k_T(t, t_l)$, and $k_M(X, X_m)$, and the regularization coefficients α_G, α_T, and $\alpha_M \in [10^{-6}, 1]$. For the regularization coefficients, the values were sampled from the logarithmic domain. The bandwidth parameter σ^2 in Eq. (4) was adjusted by $\sigma^2 = \frac{d(D)\sigma_{scale}^2}{2}$, where $d(D)$ is the median nearest distance (measured in the square Euclidean distance) of the training dataset D and σ_{scale}^2 is the bandwidth scaling parameter. The best hyperparameters were selected to be $\sigma_G = 90.079$, $\sigma_T = 98.626$, $\sigma_M = 24.894$, $\alpha_G = 1.141 \times 10^{-4}$, $\alpha_T = 3.965 \times 10^{-6}$, and $\alpha_M = 0.128$.

B. ANN model

A conventional ANN was also utilized for the EEDF prediction. The model architecture comprises of four hidden layers positioned between the input and output layers. As with the KRFD approach, the input variable X encompasses 15 spectral intensities from OES and the pressure, while the output variable Y comprises a vector of EEDF values at 1999 different energies. Each layer was constructed by a fully connected network and the ReLU activation function.[36–38] About 20% of the total data was randomly selected for the test set, and the remainder was used as the training set. By performing the fivefold cross-validation within the training set, the hyperparameters were selected using Optuna; the parameters to be selected were the number of neurons for each layer $\in \{16, \ldots, 1800\}$ and the learning rate (selected from the interval $[10^{-6}, 10^{-2}]$). While the other hyperparameters can also be optimized for a more streamlined neural network, this was not explored in this paper. The number of neurons for each hidden layer and the learning rate were optimized to be 355 neurons and 3.633×10^{-4}. Each model was trained using the stochastic gradient descent algorithm. The number of iterations (epochs) was set to 1000 for both the hyperparameter optimization and the final model training with the selected hyperparameters.

C. RF regression

The RF regression uses an ensemble of decision trees to make predictions.[39–42] Each decision tree is trained independently of the others, using a set of random samples taken from the training dataset. Every decision tree generates a prediction, with the final prediction derived from the average or weighted combination of the individual predictions. The RF regression can perform both classification for categorical data and regression for continuous variables. RF regression was used to predict a discrete scalar value of n_e. Although two RF models were created, each attached to the KRFD and ANN, they yielded almost the same results. Therefore, for the remainder of this paper, only the RF model attached to the KRFD was included. By hyperparameter optimization using Optuna, the best value for the number of trees $\in \{10, \ldots, 500\}$ and a number of features in each tree $\in \{3, \ldots, 15\}$ were selected. This optimization resulted in 89 trees and 15 features as best hyperparameter values. The branches of the tree were expanded until each leaf contained less than two samples.

IV. RESULTS: PREDICTION FROM SIMULATED DATA

This section analyzes the predictions of the trained ML models on simulated emission line intensities. We first discuss the predicted EEDFs, followed by the predicted n_e. Finally, we present the resulting intensities obtained by feeding these predicted plasma parameters back into the CRM.

A. EEDF prediction

Figure 1 compares the EEDFs predicted by KRFD and ANN with the test EEDF data. The KRFD and ANN predictions agree very well with the test data. The excellent agreement extends over more than six orders of magnitude. The small deviations observed at low energies (below about 10 eV) are likely due to the fact that the electrons in this energy range do not participate in excitation collisions of ground-state atoms. Consequently, the emission intensities are less sensitive to this region of the EEDF and, naturally, the uncertainty here is larger. Note, however, that the low-energy electrons still participate in stepwise excitations and deexcitation collisions. Therefore, the intensities are not entirely insensitive to this population of electrons and relatively reliable information about the low-energy part of the EEDF can still be obtained.

To assess the overall performance of the EEDF prediction for all test data, predicted data points were compared to their corresponding test data values. The excellent agreement is demonstrated in Fig. 2. Here, each predicted point of the EEDF is plotted against the corresponding test data for both KRFD and ANN. A good model would produce a data cloud concentrated around the $x = y$ line, with an R^2 value close to 1. The coefficient of determination R^2 is calculated by

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1} (x_i - y_i)^2}{\sum_{i=0}^{n-1} (x_i - \bar{x})^2}, \tag{7}$$

wherein x_i is a test value, y_i is a predicted value, and $\bar{x}$ is the mean of all test values. In general, both models are shown to predict the EEDF well with high R^2 values. Still, the KRFD exhibits a slightly better accuracy. Further, it is observed that the scattering around the diagonal line in Fig. 2 is the largest in the region of low EEDF values (below 10^{-6} eV^{-1}) for both methods. This corresponds to the EEDF at high electron energies (typically above the ionization energy of the argon atoms). There, the influence of the noise both in the training data as well as in the test data used for comparison
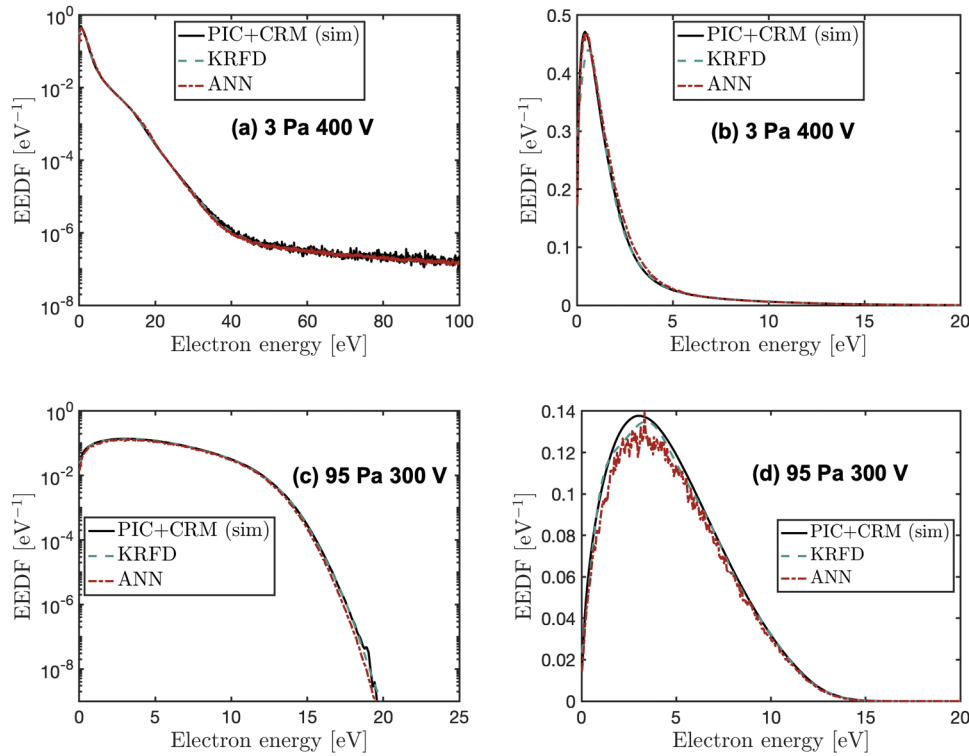
FIG. 1. EEDF predictions at 3 Pa 400 V (a) and (b) and 95 Pa 300 V (c) and (d), presented on the logarithmic scale (a) and (c) and the linear scale (b) and (d). Each subfigure shows the EEDF predicted by KRFD, by ANN, alongside the test EEDF data from simulations, labeled as "PIC+CRM (sim)."

is the largest, resulting in the observed larger deviation between predicted and test EEDF. Nevertheless, it is remarkable that both models are able to achieve good predictions on an arbitrary function even with a limited dataset for training.

B. Electron density prediction

In addition to the EEDF, ML is also used to predict n_e. Figure 3(a) shows the results from an RF model. Here, the predicted n_e values are compared against the entire simulated n_e dataset used for training and testing. Overall, the predictions from the RF model agree well with the test data, with minor deviations from the corresponding trend lines. Figure 3(b) quantifies how well

these n_e predictions agree with the test data by getting the R^2 of the predicted n_e as a function of the test n_e, as is done for the EEDF.

C. Calculating spectral intensities from predicted plasma parameters

One test for easy assessment of the effectiveness of the ML model in accurately inverting the CRM involves feeding the predicted values of the EEDF and n_e back into the CRM and checking if the model reproduces the same intensities used for predicting these input EEDF and n_e. Figure 4 shows such a comparison between the input spectra and the spectra generated by the CRM using the predicted EEDF and n_e from both the KRFD and ANN at example
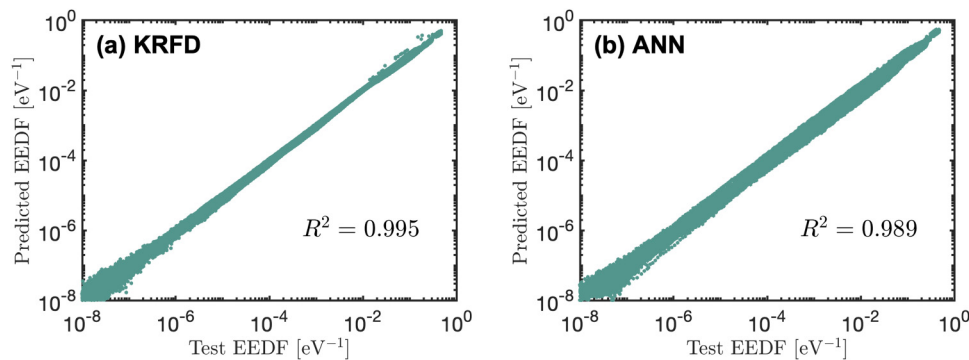

FIG. 2. Comparison of each predicted data point of the EEDF to its corresponding test data point, as predicted using KRFD (a) and ANN (b). A perfect prediction would give a diagonal line with an R^2 of 1.
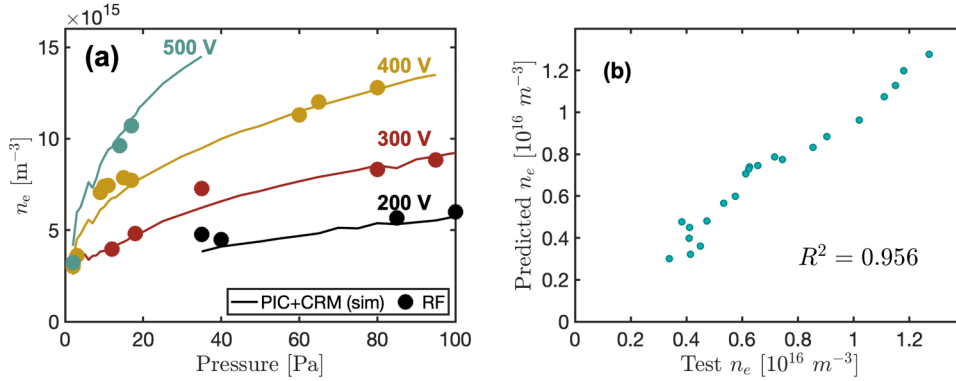
19 August 2024 04:59:59

FIG. 3. (a) Electron density n_e predicted by the RF regression (RF) for each test case, plotted alongside the entire n_e training dataset. (b) Comparison of each predicted n_e value to its corresponding test data point. A perfect prediction would give a diagonal line with R^2 of 1.

conditions (3 Pa 400 V and 95 Pa 300 V). It is seen from the figure that both the KRFD and the ANN models effectively predict plasma parameters that reproduce the test spectral data. This strong agreement highlights the models' ability to successfully invert the CRM and accurately predict the plasma parameters for test data within the training scope.

To evaluate the agreement of the predicted line intensities with the test input data, their discrepancy is quantified similar to what was done in Ref. 16. This is achieved by taking the ratio of the test input spectral intensities I_{CRM} to the intensities calculated from predictions of KRFD and ANN I_{fromML} for the wavelength values λ_k given in Table I, such that

$$r_k = I_{CRM}(\lambda_k)/I_{fromML}(\lambda_k). \tag{8}$$

The line ratios r_k are then scaled such that the average of all r_k values in a given spectrum is unity. To quantify the deviations of these r_k to 1, the root mean square is calculated as

$$RMSE = \sqrt{\frac{1}{Q}\sum_{k=1}^{Q}(r_k - 1)^2}, \tag{9}$$

where $Q = 15$ is the number of spectral lines. A lower value of the RMSE shows better agreement and an RMSE of 0 means perfect agreement between the experimental and calculated intensities.

Figure 5 shows the calculated RMSE of the predicted intensities with respect to the input test data. Relatively low RMSE values are calculated for both KRFD and ANN, with the average for KRFD generally slightly lower than that of ANN.

V. RESULTS: PREDICTION FROM EXPERIMENTAL DATA

Section IV shows the predictions of the ML models when applied to test data belonging to a similar simulation dataset. In this section, the prediction results are presented for the experimentally measured intensities using the same ML models trained with the simulation data. However, as reported in Ref. 16, it is known that the simulation data deviate from the experimental data, particularly at pressures higher than 20 Pa.

A. EEDF predictions

Similar to the simulation data, the EEDF and n_e are predicted from the experimentally measured spectra. Figures 6(a) and 6(b) present the EEDFs predicted from experimental spectra at 10 and 50 Pa, alongside the EEDF from PIC+CRM simulations for the same experimental conditions. The comparison is made for data at a peak-to-peak voltage of 300 V since experimental spectra are available only at this condition. For reference, the EEDFs under the same conditions and calculated from the PIC/MCC+DRR simulations, taken from Donko et al.,[43] (hereby denoted as Ref. 43), are also added. Compared to the PIC/MCC simulations, the PIC/MCC+DRR model
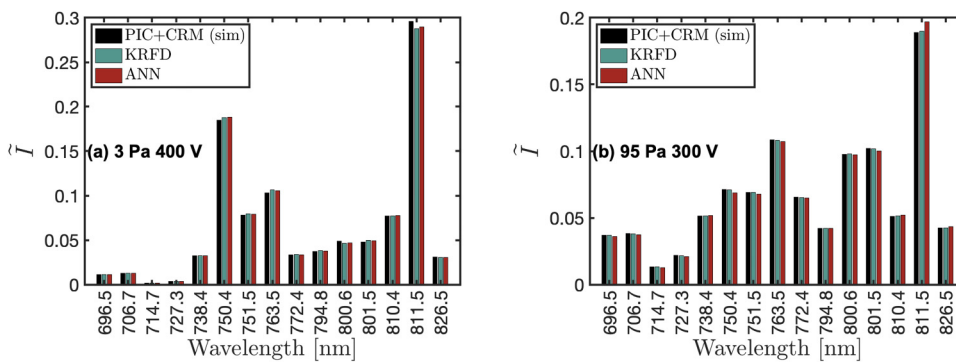

FIG. 4. Normalized spectral line intensities at 3 Pa 400 V (a) and 95 Pa 300 V (b) calculated from the predictions of KRFD and ANN, alongside the test input spectral data from simulations, labeled as "PIC+CRM (sim)."
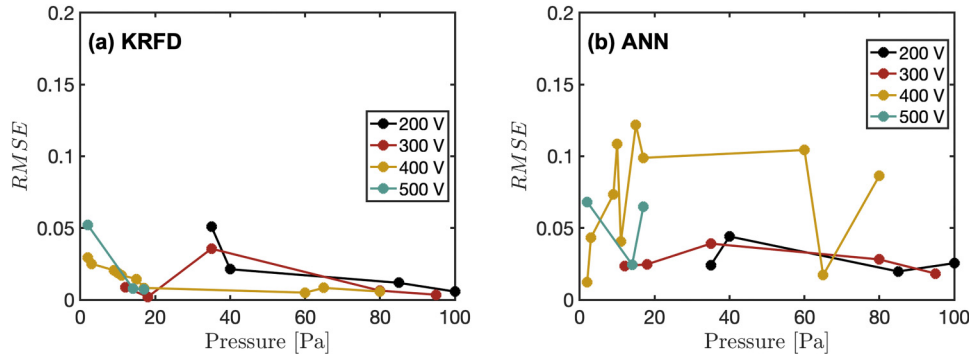
FIG. 5. Root mean square error (RMSE) of the standard deviation of the line ratios r_k [Eq. (8)] from the average ratio of 1 for the intensities calculated using the EEDF and n_e predicted by (a) KRFD and (b) ANN at different V_{pp} and pressures.

in Ref. 43 couples a basic PIC/MCC code with a diffusion–reaction–radiation (DRR) model to take into account all the processes involving the excited states. This approach demonstrates a good agreement of the calculated intensities with the experimentally measured ones.

The ANN predictions closely replicate the EEDFs from the PIC+CRM simulation [Figs. 6(a) and 6(b)]. Even at 50 Pa [Fig. 6(b)], where the simulated intensities do not agree well with the experimentally measured ones, the ANN reproduces the EEDF from the corresponding simulation well. It is possible that the ANN has learned to place a higher weight on the pressure during its training. At 10 Pa, the EEDF calculated from Ref. 43 agrees

closely with that from PIC+CRM simulations and, therefore, with the ANN predictions. On the other hand, the EEDFs predicted by KRFD do not agree well with the simulation data it is trained with. At 10 Pa, the KRFD predicted an EEDF that deviates from both simulation results. Moreover, the EEDFs appear "bumpy" in contrast to the ANN predictions. This is likely an artefact of the chosen kernel function [Eq. (4)]. At 50 Pa, however, the KRFD predictions agree closer to that from Ref. 43 than that from PIC+CRM simulations. This is an interesting effect, considering that the training of the KRFD involved EEDFs with a rather different shape at these conditions [see the PIC+CRM curve in Fig. 6(b)].
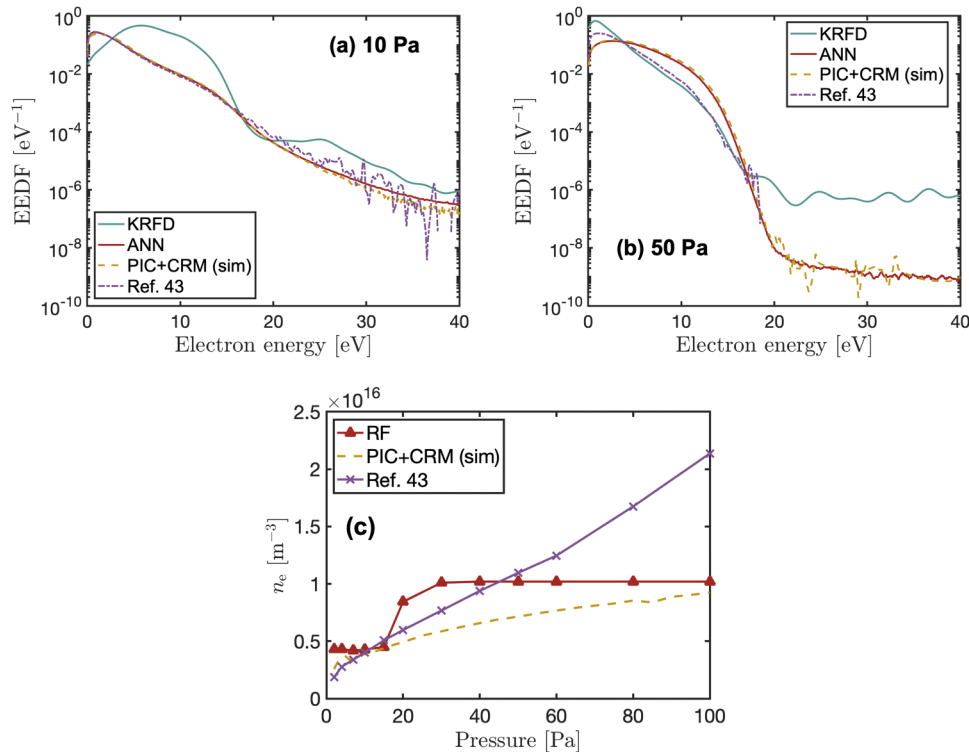

FIG. 6. Prediction results from experimentally measured input data: predicted EEDFs using KRFD and ANN at 10 and 50 Pa, plotted alongside the EEDF from PIC+CRM simulations and from simulations by Donko et al. (Ref. 43) at (a) 10 Pa, 300 V and (b) 50 Pa, 300 V. (c) n_e predicted using RF, along with the corresponding n_e from PIC+CRM simulations and from simulations by Donko et al. (Ref. 43) at 300 V.

B. Electron density predictions

Similar to the simulation data, an RF model was used to predict n_e from experimentally measured spectra. Figure 6(c) shows n_e values predicted by RF, along with the training data and n_e obtained in Ref. 43. In this case, we do not observe the same pressure dependence of the predicted n_e as seen in the PIC+CRM or that from Ref. 43. Instead, the predicted n_e values seem to fall into two distinct regions: low pressure (below 20 Pa) and high pressure (above 20 Pa). In each region, n_e values remain nearly constant. It was already demonstrated that the developed ML models provide a good inverse of the CRM. The unexpected behavior of n_e predicted from experimental spectra likely hints toward the fact that the CRM itself might not provide a good description of the actual processes in the experiment and would require an extension. However, this goes beyond the scope of the current work.

C. Calculating spectral intensities from predicted plasma parameters

The prediction for the EEDF and n_e from the experimental intensities is fed back into the CRM to calculate the corresponding intensities. This is done to investigate to what extent the inversion of the CRM by the ANN and the KRFD works with experimental spectra. An example of these results is shown in Fig. 7, that is for the 10 Pa case. As is the case for the predicted EEDF, the intensities from the ANN predictions yield intensities that closely agree with the PIC+CRM simulations. Given that the simulations show better agreement with experiments at low pressures, the ANN is also able to yield good results at low pressures.
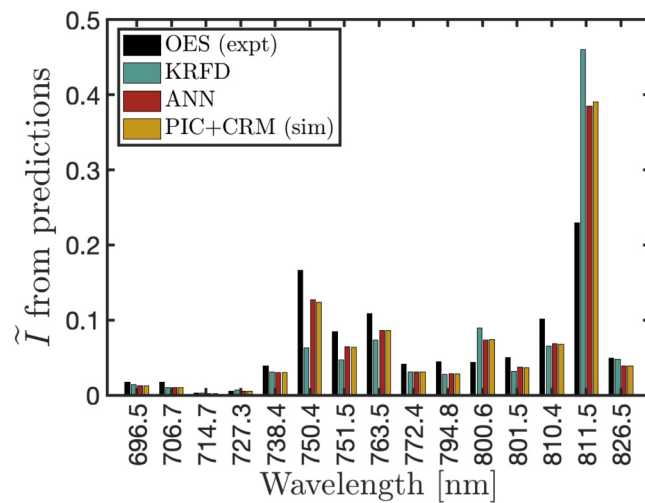
The close agreement between the ANN predictions and the simulation spectra might seem surprising at first, given that for pressures below 20 Pa, there is a discrepancy of about a factor of 2 between the electron densities used for generating these spectral data [Fig. 6(c)]. However, as demonstrated in Ref. 16, the emission intensities calculated by this CRM have a strong linear dependence on n_e and a much weaker nonlinear contribution. Then, using a normalization of the emission intensities to their sum removes the strong linear dependence on n_e, leaving only the much weaker nonlinear contribution, which makes for the slight differences in the intensities by ANN and PIC+CRM seen in Fig. 7.

The intensities obtained from the KRFD predictions at 10 Pa significantly deviate from the experimental values [Fig. 6(a)]. This is likely due to the non-negligible differences in the EEDF predicted by KRFD from the actual EEDF [Fig. 6(a)]. Particularly, the deviations in the region of electron energies close to the excitation thresholds for Ar (11–15 eV) are probably the main cause for the discrepancies observed in the spectral intensities.

In the pressure range above 20 Pa, it is known[16] that the intensities calculated by PIC+CRM deviate from the experimental measurements. It is, thus, of relevance to investigate how the intensities calculated from the CRM with the EEDF and n_e predicted by the ANN and KRFD from the experimental intensities compare to the actual experimental spectra, i.e., to what extent the ANN and the KRFD can invert the CRM for spectra from the experiment. Such a comparison is plotted in Fig. 8 for the 50 Pa case. Here, similar observations can be made about ANN as before: it replicates the PIC+CRM results. Notably, it does so even with the larger difference in n_e in this pressure range [Fig. 6(c)]. This is again likely an artefact of the normalization of the intensities.

FIG. 7. Normalized intensities at 10 Pa and 300 V peak-to-peak voltage calculated from the predictions of KRFD and ANN obtained by feeding the experimental spectra. For comparison, the underlying experimentally measured intensities denoted as "OES (expt)" together with the spectrum from simulations "PIC+CRM (sim)" with the same conditions are also presented. The latter data were also part of the training dataset.
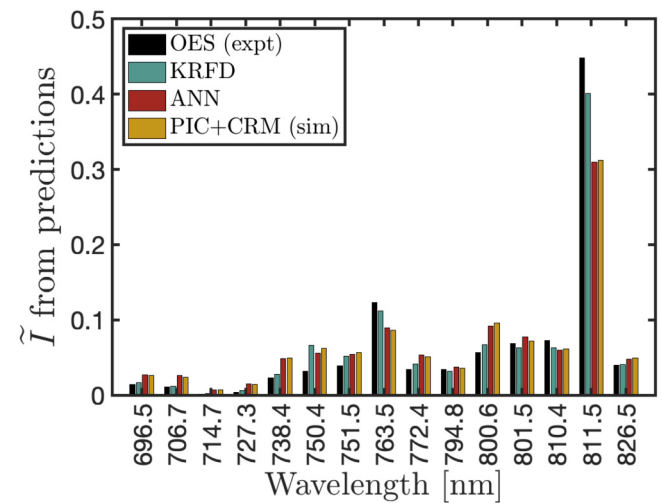
FIG. 8. Normalized intensities at 50 Pa and 300 V peak-to-peak voltage calculated from the predictions of KRFD and ANN obtained by feeding the experimental spectra. For comparison, the underlying experimentally measured intensities denoted as "OES (expt)" together with the spectrum from simulations "PIC+CRM (sim)" with the same conditions are also presented. The latter data were also part of the training dataset.
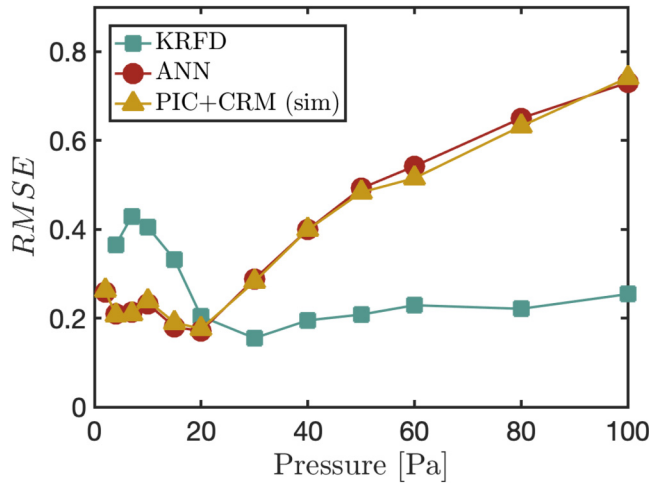
FIG. 9. Root mean square error (RMSE) of the ratios r_k of the intensities from "PIC+CRM" and from KRFD and ANN predictions to the experimentally measured intensities at a peak-to-peak voltage of 300 V.

FIG. 10. Ar $1s_5$ metastable density as measured by TDLAS,[16] simulated using PIC+CRM,[16] and calculated from the predictions from KRFD and ANN via the CRM at a peak-to-peak voltage of 300 V.

However, since at higher pressures, the PIC+CRM data deviate from the experimental data, the resulting intensities from ANN do not agree with the experimental one at 50 Pa. Surprisingly, the KRFD predictions result in intensities that are in better agreement with the experimental intensities. This is likely related to the better agreement of the KRFD prediction for the EEDF with the more realistic simulations of Ref. 43.

The differences in the calculated spectra from the experimentally measured ones is quantified through the RMSE [Eq. (9)]. The resulting RMSE for KRFD, ANN, and the PIC+CRM case are shown in Fig. 9. As expected, the ANN results closely follow the trend of PIC+CRM, while KRFD shows a different behavior. It shows poorer agreement (a higher RMSE) at low pressures and better consistency (a relatively lower RMSE) with the experimental spectra at higher pressures (above about 20 Pa). At the lowest pressure (2 Pa), the EEDF predicted by KRFD did not allow the CRM code to converge, and a data point is thus missing for this case.

In addition to the intensities, the density of the lowest excited state of argon, $1s_5$, is also calculated from the CRM. In Ref. 16, $1s_5$ density from the PIC + CRM simulations and the TDLAS experiments are compared. Figure 10 shows the same TDLAS data, which is given along with the resulting $1s_5$ density obtained from PIC+CRM simulations and those predicted using plasma parameters from ML. Unsurprisingly, the ANN and the PIC+CRM values are very close. At low pressures, they also follow the TDLAS data. However, at high pressures (above 20 Pa), the predicted $1s_5$ density from ANN continuously increases and deviates from the experimental measurements. Conversely, the KRFD predictions show poor agreement with the experiment at low pressures but surprisingly good agreement at high pressures.

To use the ML model for other experimental setups, the spectrometer used for measurements must be calibrated according to its wavelength-dependent sensitivity and be able to distinguish the specific lines of interest. The ML model is trained using CRM data
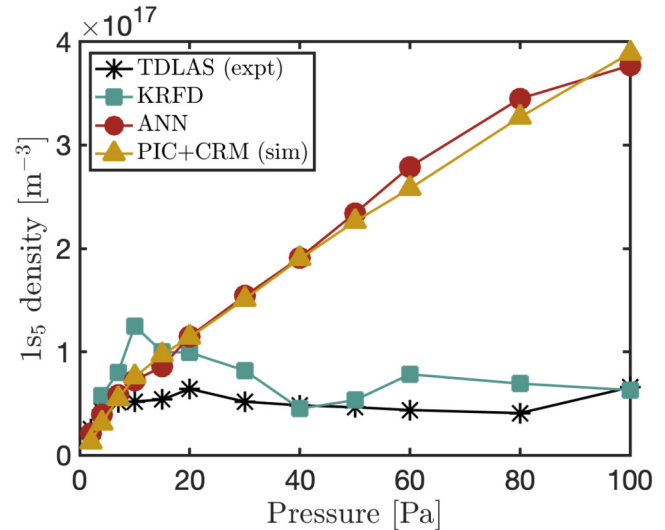
specific to an argon plasma with pressures ranging from 2 to 100 Pa, which sets the applicable experimental range of the model. Moreover, the metastable diffusion profile and self-absorption profile of the excited states assumed in the CRM also depend on the electrode radius and on the length of the gap between the electrodes. Therefore, this current model is limited to plasma sources having a geometry close to the one used in the experimental setup in this study, i.e., an electrode radius of ≈ 7 cm and gap ≈ 4 cm.

VI. CONCLUSIONS

In this study, ML models were developed to predict the normalized EEDF and n_e of Ar plasma from its relative (i.e., normalized) optical emission line intensities at various gas pressures. The models were trained on 86 sets of numerical simulation data obtained from the 1D PIC/MCC simulation code coupled with a CRM code described in Ref. 16 for Ar CCPs under various discharge conditions. Two different types of models, i.e., KRFD and an ANN, were used to predict the normalized EEDF and the RF regression was used to predict n_e. The input data for the ML models are the gas pressure and the relative emission line intensities. When the simulated OES data were used for the input data, the ML models predicted the corresponding EEDF and n_e with high accuracy. In other words, the ML models developed in this study performed well for what they were trained for.

The simulation data of Ref. 16 are known to agree well with experimental data at a gas pressure of up to about 20 Pa. However, even under these conditions, some discrepancy always exists between the simulated and experimentally measured OES data. The discrepancy stems from various reasons, ranging from the inaccuracy of the CRM to the profile effects of the three-dimensional actual plasma.

If the developed ML models are robust enough, they may also work for the experimentally measured OES data used as their input data for a gas pressure of up to about 20 Pa. The robustness of the ML models developed in this study was then tested. It was found that the ANN-based ML model was able to predict the normalized EEDF reasonably well at a pressure of 20 Pa or lower for the experimentally obtained relative line intensities and the gas pressure. However, under the same conditions, the predictions by the KRFD-based ML model exhibited relatively large errors. The RF regression model for n_e was not able to show the pressure dependence under the same conditions and, therefore, deemed not satisfactory. The results show that, in general, the ML models, which were trained only with the simulation data and performed well with simulated OES data, were not robust enough to handle the discrepancy between the simulated and experimental OES data.

Therefore, the future direction of this study includes the improvement of the robustness of the ML models. The use of a larger and more diverse set of simulation data and experimental data can also be used to improve the model. However, a fundamental question remains as to whether the relative emission line intensities contain sufficient information on the electron density and its energy distribution. Although the optical emission's relative line intensity spectrum may depend on n_e, the dependence may be too subtle to be detected if other effects modify the spectrum shape. Furthermore, the emission spectra are likely to be more sensitive to a particular energy range of the EEDF. Therefore, different EEDFs can give rise to similar OES spectra. These issues will also be addressed in future work.

The recent study of Ref. 43 performed 1D PIC/MCC simulations with stepwise excitation and ionization and a self-consistent CRM, which produced OES line intensity data in good agreement with the experimental data for Ar CCPs at a gas pressure up to 100 Pa. Training of the ML models presented here with large-scale simulation data generated by the simulation code of Ref. 43 can make these models valid in all pressure ranges.

Using the EEDF presented in Ref. 43, we also tested the current ML models of this study at a high pressure of up to 100 Pa with experimentally obtained relative optical emission line intensities. As discussed earlier, the current ML models were not trained for the actual plasma at a pressure higher than 20 Pa (where the training simulation data do not agree well with the experimental data) Therefore, it is not surprising that most predictions failed under these conditions. However, the KRFD-based ML model predicted the EEDF and other plasma properties similar to those data obtained from the latest numerical simulations given in Ref. 43 at a pressure above 20 Pa. The KRFD-based ML model was not trained for such discharges and, therefore, these predictions should not be trusted. Nevertheless, the KRFD-based ML model's seemingly good performance at high pressure reflects some nature of optical emission spectra of Ar discharges and its discussion is deferred to a future study.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Fatima Jenina Arellano: Conceptualization (equal); Data curation (equal); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (equal); Software (equal); Validation (equal); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **Minoru Kusaba:** Formal analysis (equal); Methodology (equal); Software (equal); Supervision (equal); Validation (equal); Writing – review & editing (equal). **Stephen Wu:** Formal analysis (equal); Methodology (equal); Project administration (equal); Software (equal); Supervision (equal). **Ryo Yoshida:** Formal analysis (equal); Methodology (equal); Project administration (equal); Software (equal); Supervision (equal); Writing – review & editing (equal). **Zoltán Donkó:** Conceptualization (equal); Data curation (lead); Funding acquisition (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Writing – review & editing (equal). **Peter Hartmann:** Data curation (equal); Methodology (equal); Resources (equal); Software (equal). **Tsanko V. Tsankov:** Data curation (equal); Formal analysis (equal); Software (equal); Supervision (equal); Writing – review & editing (equal). **Satoshi Hamaguchi:** Conceptualization (lead); Data curation (equal); Formal analysis (equal); Funding acquisition (lead); Methodology (equal); Project administration (equal); Resources (equal); Supervision (lead); Visualization (equal); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]G. S. Oehrlein and S. Hamaguchi, Plasma Sources Sci. Technol. **27**, 023001 (2018).
[2]K.-D. Weltmann et al., Plasma Process. Polym. **16**, 1800118 (2019).
[3]I. Adamovich et al. J. Phys. D: Appl. Phys. **55**, 373001 (2022).
[4]R. Engeln, B. Klarenaar, and O. Guaitella, Plasma Sources Sci. Technol. **29**, 063001 (2020).
[5]J. Benedikt, H. Kersten, and A. Piel, Plasma Sources Sci. Technol. **30**, 033001 (2021).

[6] B. M. Goldberg, T. Hoder, and R. Brandenburg, Plasma Sources Sci. Technol. **31**, 073001 (2022).

[7] Ts. V. Tsankov and U. Czarnetzki, Plasma Sources Sci. Technol. **26**, 055003 (2017).

[8] G. Gifford, "Applications of optical emission spectroscopy in plasma manufacturing systems," in *Advanced Techniques for Integrated Circuit Processing* (International Society for Optics and Photonics, Santa Clara, CA, 1991), Vol. 1392, pp. 454–465.

[9] R. Yang and R. Chen, Sensors **10**, 5703 (2010).

[10] V. M. Donnelly, J. Phys. D: Appl. Phys. **37**, R217 (2004).

[11] E. Restrepo and A. Devia, J. Vac. Sci. Technol. A **22**, 377 (2004).

[12] N. Ohno, M. A. Razzak, H. Ukai, S. Takamura, and Y. Uesugi, Plasma Fusion Res. **1**, 028 (2006).

[13] D. Hope, T. Cox, and V. Deshmukh, Vacuum **37**, 275 (1987).

[14] T. Mehdi, P. Legrand, J. Dauchot, M. Wautelet, and M. Hecq, Spectrochim. Acta B: At. Spectrosc. **48**, 1023 (1993).

[15] X.-M. Zhu and Y.-K. Pu, J. Phys. D: Appl. Phys. **43**, 015204 (2009).

[16] F. J. Arellano, M. Gyulai, Z. Donkó, P. Hartmann, Ts. V. Tsankov, U. Czarnetzki, and S. Hamaguchi, Plasma Sources Sci. Technol. **32**, 125007 (2023).

[17] D. Lopaev, A. Volynets, S. Zyryanov, A. Zotovich, and A. Rakhimov, J. Phys. D: Appl. Phys. **50**, 075202 (2017).

[18] S. Iordanova and I. Koleva, Spectrochim. Acta B: At. Spectrosc. **62**, 344 (2007).

[19] A. Palmero, E. Van Hattum, H. Rudolph, and F. Habraken, J. Appl. Phys. **101**, 053306 (2007).

[20] Z. Navrátil, P. Dvořák, O. Brzobohatý, and D. Trunec, J. Phys. D: Appl. Phys. **43**, 505203 (2010).

[21] S. Siepa, S. Danko, Ts. V. Tsankov, T. Mussenbrock, and U. Czarnetzki, J. Phys. D: Appl. Phys. **47**, 445201 (2014).

[22] K. Evdokimov, M. Konischev, V. Pichugin, and Z. Sun, Resour.-Effic. Technol. **3**, 187 (2017).

[23] E. Desjardins, M. Laurent, A. Durocher-Jean, G. Laroche, N. Gherardi, N. Naudé, and L. Stafford, Plasma Sources Sci. Technol. **27**, 015015 (2018).

[24] K.-B. Chai and D.-H. Kwon, Spectrochim. Acta B: At. Spectrosc. **183**, 106269 (2021).

[25] H. Horita, D. Kuwahara, H. Akatsuka, and S. Shinohara, AIP Adv. **11**, 075226 (2021).

[26] D. Nishijima, S. Kajita, and G. Tynan, Rev. Sci. Instrum. **92**, 023505 (2021).

[27] J.-H. Park, J.-H. Cho, J.-S. Yoon, and J.-H. Song, Coatings **11**, 1221 (2021).

[28] K. Shojaei and L. Mangolini, J. Phys. D: Appl. Phys. **54**, 265202 (2021).

[29] T. van der Gaag, H. Onishi, and H. Akatsuka, Phys. Plasmas **28**, 033511 (2021).

[30] Y. Ralchenko, Mem. Soc. Astron. Ital. Supplement **8**, 96 (2005).

[31] U. Fantz, Plasma Sources Sci. Technol. **15**, S137 (2006).

[32] Z. Donkó, Plasma Sources Sci. Technol. **20**, 024001 (2011).

[33] Z. Donkó, A. Derzsi, M. Vass, B. Horváth, S. Wilczek, B. Hartmann, and P. Hartmann, Plasma Sources Sci. Technol. **30**, 095017 (2021).

[34] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, August 2019 (Association for Computing Machinery, New York, 2019), pp. 2623–2631.

[35] M. Iwayama, S. Wu, C. Liu, and R. Yoshida, J. Chem. Inf. Model. **62**, 4837 (2022).

[36] T. Kessler, G. Dorian, and J. H. Mack, "Application of a rectified linear unit (ReLU) based artificial neural network to cetane number predictions," in *Internal Combustion Engine Division Fall Technical Conference*, Vol. 58318 (American Society of Mechanical Engineers, Seattle, WA, 2017), p. V001T02A006.

[37] J. E. Choi and S. J. Hong, Meas.: Sens. **16**, 100046 (2021).

[38] A. D. Bonzanini, K. Shao, D. B. Graves, S. Hamaguchi, and A. Mesbah, Plasma Sources Sci. Technol. **32**, 024003 (2023).

[39] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1 (IEEE, Montreal, 1995), pp. 278–282.

[40] L. Breiman, Mach. Learn. **45**, 5 (2001).

[41] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2009), Vol. 2.

[42] F. Pedregosa *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011).

[43] Z. Donkó, Ts. V. Tsankov, P. Hartmann, F. J. Arellano, U. Czarnetzki, and S. Hamaguchi, "Self-consistent calculation of the optical emission spectrum of an argon capacitively coupled plasma based on the coupling of particle simulation with a collisional-radiative model," J. Phys. D Appl. Phys. **57**, 375209 (2024).