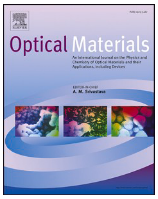
Optical Materials 174 (2026) 117917

Contents lists available at ScienceDirect

Optical Materials

journal homepage: www.elsevier.com/locate/optmat



Research Article

Predicting fs-laser-induced NV centers with PCA and neural networks

Murilo Neco Saraiva, Orlando Marbello Ospina, Lucas Konaka Nolasco, Renan Souza Cunha, Lucas Nunes Sales de Andrade, Sergio Ricardo Muniz, Cleber Renato Mendonca [*]

Instituto de Física de São Carlos, Universidade de São Paulo, CP 369, São Carlos, SP, 13560-970, Brazil

ABSTRACT

Diamond hosts a variety of lattice defects, among which nitrogen-vacancy (NV) centers stand out due to their relevance in quantum photonics with optically addressable qubits. Yet, the complex laser–material interactions governing its formation are not fully understood, and the influence of laser parameters on NV generation still raises open questions. Here, we investigate the generation of NV centers using principal component analysis (PCA) and artificial neural networks (ANNs) as predictive tools based on femtosecond laser parameters. Experimental results from femtosecond laser micromachining of diamond provided the dataset for our analysis. We employed PCA to reduce data dimensionality and uncover dominant experimental trends, while a multilayer perceptron model was trained to predict NV center generation under simulated conditions. GridSearch optimization and Leave-One-Out cross-validation (LOOCV) ensured the best performance and robustness of the ANN. Our results reveal that NV center generation is directly proportional to laser peak fluence and inversely proportional to pulse duration and excitation wavelength. Notably, PCA and ANN modeling independently converged on consistent trends, reinforcing the reliability of the observed parameter–defect relationships. This convergence supports the development of predictive frameworks for controlled color center generation in diamond with greater precision.

1. Introduction

Machine learning (ML) is a branch of computer science that explores the ability of algorithms to learn and solve problems by mimicking human intelligence [1]. Artificial neural networks (ANNs) are ML models inspired by the architecture of biological neural systems, created by simulating a network of artificial neurons. The way that ANNs “learn” depends on the input data and can generally be classified into supervised, unsupervised, and semi-supervised learning [2]. In supervised learning, the ANN adjusts its internal weights based on labeled examples to minimize the error between predicted and actual outputs [3], while unsupervised ANNs, such as autoencoders, learn data patterns without predefined labels [4]. Semi-supervised models combine both labeled and unlabeled data to improve generalization in scenarios with limited experimental information [5]. In this work, we employed a supervised ANN model known as a multilayer perceptron. Its training process relies on a backpropagation algorithm combined with an optimization method to iteratively minimize a loss function, typically defined as the sum of squared differences between predictions and targets [6–8]. These properties make ANNs highly versatile and particularly suited for modeling systems where multiple interdependent variables influence the outcome.

Given their ability to handle complex tasks, ANNs have been applied across a wide range of domains. In photonics, they are effective tools for accelerating the intricate design process of advanced photonic devices and structures [9–11]. Previous studies have used ANNs to predict optimal parameters, for instance, in laser microdrilling of titanium nitride–alumina composites [12] and to model the relationship between laser micromachining parameters and quality outcomes for AISI H13 hardened tool steel [13]. Additionally, they have been used to predict the optimal machining parameters for generating the maximum groove depth in tungsten-molybdenum high-speed steel [14], assist in the 3D printing of microneedle-based devices [15], and help predict the wettability of microtextured surfaces [16]. These studies highlight the strength of ANNs in modeling and optimizing complex laser-based fabrication processes, offering faster, cheaper, and more scalable access to optimal conditions than would be feasible through experiments alone.

Nitrogen-vacancy (NV) color center defects consist of a substitutional nitrogen atom associated with a vacancy in a nearby lattice site in diamond. Their long spin coherence times, combined with optical addressability enabled by the interplay between optical and magnetic resonance techniques, make the negatively charged NV^- centers strong candidates for room-temperature quantum technologies [17–23]. To this end, femtosecond laser micromachining offers a promising route for the localized generation of such defects, which is an essential step toward scalable quantum applications. However, the process remains challenging to control, as defect formation depends on a combination of

[*] Corresponding author.
E-mail address: crmendon@ifsc.usp.br (C.R. Mendonca).

https://doi.org/10.1016/j.optmat.2026.117917
Received 20 August 2025; Received in revised form 9 January 2026; Accepted 21 January 2026
Available online 23 January 2026

sample-specific aspects and interrelated laser parameters, such as peak fluence, wavelength, and pulse duration. The sensitivity of the outcome to slight variations, along with the high dimensionality of the parameter space, makes it challenging to anticipate results or define optimal fabrication parameters solely through experimental exploration.

Here, we applied ANNs to predict the generation of NV centers in diamond given specific microfabrication parameters. A Yb:KGW femtosecond laser system was used to generate NV color centers in a CVD diamond sample, providing the dataset for our computational analysis. As a preliminary step before regression analysis, principal component analysis (PCA) was performed to analyze the dataset structure, identify outliers, and reveal major trends among the variables. PCA offers a more streamlined approach than traditional exploratory methods because it transforms correlated variables into orthogonal components that explain most of the variance. This dimensionality reduction enables a quantitative visualization of key trends in 2–3 dimensions, which is particularly helpful when there are many predictors relative to the sample size. Additionally, it reduces noise and the risk of overfitting, as the first few principal components capture the primary trends.

A multilayer perceptron regression model was optimized and trained to predict the generation of laser-induced NV centers using the fabrication parameters as inputs. Leave-One-Out cross-validation (LOOCV) was employed to ensure robust generalization of the ML algorithm. These tools independently demonstrated that NV center generation increases with laser peak fluence and decreases with both pulse duration and excitation wavelength. The combination of PCA for dimensionality reduction and regression-based ANN proved to be a compelling methodology for predictive modeling in laser–material interactions, reducing the need for experimental trial-and-error approaches in identifying optimal laser configuration.

2. Methodology

2.1. Dataset

We used a dataset comprising 75 experimental observations derived from laser-generated NV centers in diamond. NV color centers were produced via fs-laser processing of a CVD diamond sample purchased from Element Six, with nitrogen and boron impurity of 0.1 and 0.05 ppm, respectively. The method relies on the generation of vacancies as a result of the fs-laser interaction, which can lead to the NV centers formation these vacancies are located near nitrogen impurities [18,24,25]. Such processing was performed using a Yb:KGW femtosecond laser system emitting pulses at 1030, 515, or 343 nm. A Pockels cell-based pulse selector controlled the repetition rate from 100 Hz to 1 MHz and a built-in stretcher-compressor unit tuned the pulse duration between 185 fs and 1 ps. A 40 × /0.65 NA microscope objective focused the beam on the sample, which was mounted on a computer-controlled three-axis motorized translation stage.

Areas of approximately 100 × 100 μm^2 were laser-processed, under different conditions, on the surface of the diamond sample using a scanning speed of 10 μm/s. For each wavelength, 343, 515, and 1030 nm, the pulse duration ranged from 185 fs to 1 ps. The peak fluence was varied from the minimum value required to induce any modification in the material up to the maximum value preceding the prominent onset of graphitization, that is, before the accumulation of residual lattice damage. This procedure resulted in 75 distinct structures. After irradiation, the sample was annealed at 680 °C to promote vacancy diffusion (thereby increasing the probability of NV center generation) and remove most of the amorphous carbon generated by the ablation process. Chemical cleaning was subsequently performed using an acid mixture to remove residual impurities from the CVD diamond processing. NV centers were identified using confocal microscopy (Zeiss model LSM-780) with 543 nm laser excitation, collecting the emission at 621–700 nm to capture NV^- fluorescence, whose zero-phonon line lies at 637 nm [26–28], and via ODMR (Optically Detected Magnetic Resonance) measurements. Fig. 1 displays typical zero-field ODMR results, where the microwave frequency was swept around the 2.8 GHz values with a 1 MHz step. The decrease in the signal around the 2850 MHz frequency further confirms the presence of the NV^- in the microstructures – a characteristical NV^- behavior [29–31]. We quantified NV^- formation by measuring the bright red emission area as a percentage of the irradiated region. More specifications regarding the NV center generation can be found in Ref. [32]. Fig. 2 shows the micromachined structures obtained at different laser fluences.

2.2. Principal components analysis

We employed PCA as an unsupervised data exploration to reveal underlying patterns and relationships within the dataset. The primary objectives were to reduce dimensionality while retaining the most relevant variance, visualize the data distribution in a lower-dimensional space, and identify potential clusters and outliers [33,34]. Unlike predictive modeling approaches, PCA does not infer causal relationships; instead, it provides a comprehensive overview of data structure, aiding in the interpretation of experimental trends.

Principal component analysis is most effective when variables exhibit statistical dependence, as it seeks to transform correlated variables into a set of orthogonal principal components [33,35]. The correlation matrix, shown in Fig. 3, displays the Pearson correlation coefficients between all variables of the original dataset.

Bartlett's sphericity test was used to evaluate the suitability of the dataset for PCA by testing whether the correlation matrix significantly deviates from an identity matrix of the same dimension [36]. In this case, a p-value of 4.4×10^{-46} confirmed the viability of the dataset for PCA application. Before building the PCA model, we standardized the data using the Z-score transformation, as defined by

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \tag{1}$$

in which x_{ij} is the original value of the j-th variable for the i-th observation, $\overline{x}_j$ is the mean of the j-th variable across all observations, and s_j is the standard deviation of the j-th variable. Each value was adjusted by subtracting the mean and then scaling by the standard deviation [37]. This step is necessary to prevent the PCA model from being disproportionately influenced by variables with larger numerical ranges.

We applied a cumulative variance threshold of 80 % to determine the number of principal components to use in the PCA model. This threshold represents a balance between capturing sufficient information and maintaining model simplicity. It serves as a reliable retention criterion

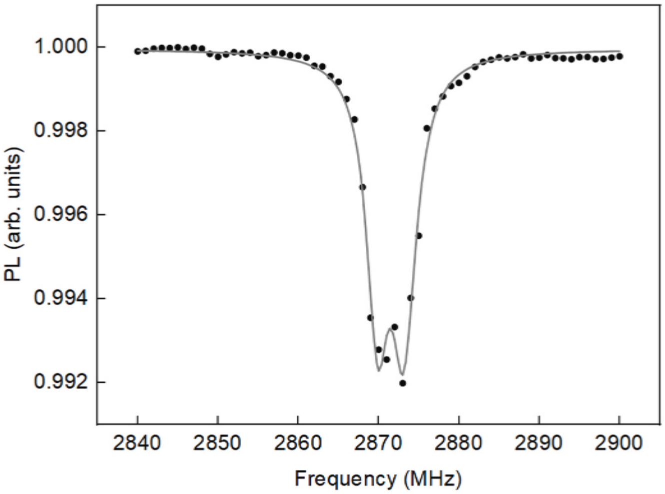


Fig. 1. – Zero-field ODMR measurement of the micromachined area at 515 nm.

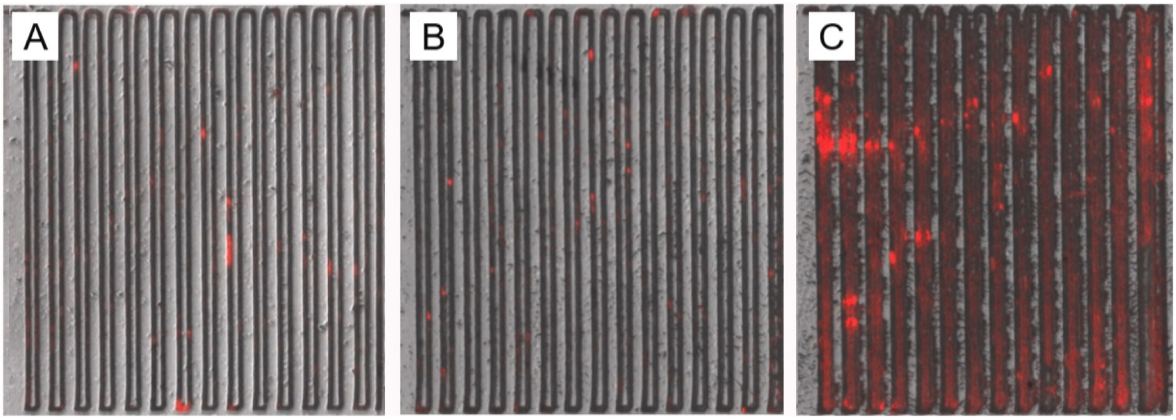


Fig. 2. – Confocal microscopy images of the micromachined structures on the surface of a CVD diamond sample. Micromachining was performed at 515 nm with 1000 fs pulses, using fluences of 0.79 J/cm^2 (A), 1.05 J/cm^2 (B), and 1.26 J/cm^2 (C).

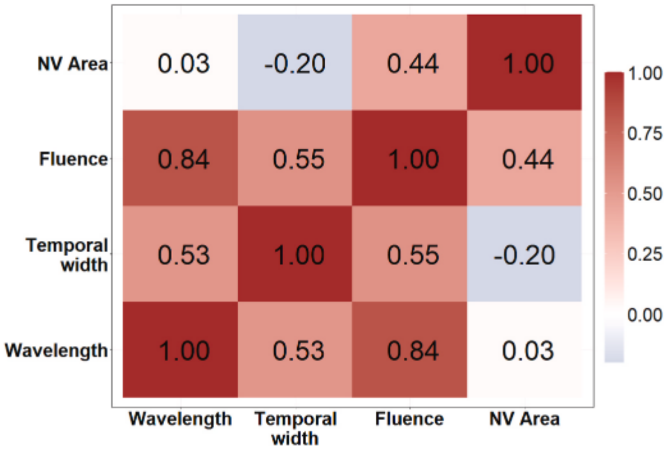


Fig. 3. Correlation matrix showing the pairwise relationships between the experimental parameters used as input for PCA. Notably, fluence and wavelength are strongly correlated (r = 0.84), while temporal width exhibits low correlation with other variables.

to ensure the model's robustness in terms of cumulative variance, considering the sample size. The PCA results were then analyzed using score and loading plots. The score plot visualizes the distribution of the new data points along the principal component axes, enabling the identification of patterns and clusters. In contrast, the loading plot represents the contribution of each original variable to the principal components, highlighting the correlation between variables. Specifically, an angle θ < 90° indicates a positive correlation, θ ~90° suggests that the variables are independent, and θ > 90° signifies a negative correlation between variables [34,38].

2.3. Artificial neural network

As part of our ANN–based regression framework, we adopted a supervised learning approach using the MLPRegressor (MLP) from the scikit-learn library [39]. Laser fluence, wavelength, and pulse duration were the predictor variables used by the MLP to predict the generation of NV centers. Table 1 illustrates a portion of the experimental dataset used to train and test the MLP model. We applied the StandardScaler to the input variables to normalize the data and improve the convergence of the model [40,41]. This pre-processing is an essential step, as it speeds up computational calculations and improves accuracy [42,43]. StandardScaler standardized the data by using Eq. (1) [44]. No additional data transformation techniques were applied, and the original variables remained unchanged. Following standardization, the hyperparameters of the MLP were tuned to refine its performance further.

Table 1
– Reduced experimental dataset. Behavior of NV center area (%) as a function of wavelength (nm), pulse duration (fs), and fluence (J/cm^2).

Wavelength (nm)	Pulse duration (fs)	Fluence (J/cm^2)	NV area (%)
1030	216	0.47	0.5
1030	216	0.56	0.3
515	185	0.14	0.2
515	185	0.2	0.4
343	185	0.16	0.9
343	185	0.22	1.4
343	500	0.37	1.8

The performance of artificial neural networks is strongly influenced by the configuration of user-defined parameters [45,46]. They have several adjustable hyperparameters that shape their architecture and influence their effectiveness [47]. To identify the optimal configuration for the MLP model, tuning was carried out using GridSearch in a train set, which involves an exhaustive search within predefined limits for each hyperparameter of the ANN. For this purpose, the dataset was randomly split into 75 % for training and 25 % for testing, using the train_test_split function with the random_state parameter set to 42 to ensure reproducibility. Leave-One-Out (LOOCV) cross-validation was used during hyperparameter tuning process to evaluate the model's generalization capability. In LOOCV, each iteration selects a single sample as the test set, while the remaining N-1 samples are used for training, repeating this procedure for all N samples in the dataset. This approach is particularly suitable for small datasets, as it maximizes the use of available data and provides an unbiased estimate of model performance [48,49]. The tuning process focused on optimizing the size of the hidden layers, the activation function, the optimization algorithm and the regularization parameter.

The configuration that yielded the best predictive performance in the tuning process consisted of two hidden layers with 21 and 10 neurons, respectively. The model employed the relu activation function, and the 'lbfgs' solver, a quasi-Newton optimization algorithm well-suited for small datasets. A regularization strength of 0.1, initial learning rate of 0.01, and a fixed random seed of 42 were used. The performance of the best model was evaluated using the lowest mean absolute percentage error (MAPE), a widely adopted metric known for its interpretability and relevance in regression tasks [50,51]. These hyperparameters were then used to train a final model on the entire training set, which was subsequently used to make predictions on the external test set. The global predictive performance of the ANN was assessed using four commonly used regression metrics: mean squared error (MSE), mean absolute error (MAE), MAPE, and the coefficient of determination (R^2). All implementations were carried out in Python using the scikit-learn, NumPy, and pandas libraries.

3. Results and discussion

3.1. Unsupervised ML analysis

Fig. 4 shows the cumulative variance for each principal component (PC), indicating that the first two principal components account for 88 % of the data's variability, with PC1 contributing the largest share at 58 %.

The loading plot, also known as cos2(θ) plot, in Fig. 5 shows the variable distribution in the new principal components space. Specifically, the loading represents the Pearson correlation between the principal components used and the original variables as given by

$$\cos^2(v_j, PC_k) = \frac{(l_{jk})^2}{\sum_{n=1}^{p} (l_{jn})^2}, \tag{2}$$

where l_{jk} represents the loading value for the j-th variable in the k-th principal component and $\sum_{n=1}^{p} (l_{jn})^2$ the sum of the square of all the variable loadings. Therefore, the greater the component load, the more influence the variable has on that component [52,53]. In this case, the graph can be used to confirm that the variables are well represented by PC1 and PC2 and that no information is being lost in the other components.

Fig. 5(A) demonstrates that the first two principal components effectively capture the variance of all variables, whereas Fig. 5(B) indicates that higher order components contribute negligibly. These results support the dimensionality reduction approach with minimal information loss.

The NV area is mainly described by PC2. Based on its relative angular orientation in the component space, a positive correlation is observed with the ablation threshold fluence and the component, while a negative correlation exists with the pulse duration. Additionally, there is no correlation with the excitation wavelength. These findings suggest that maximizing the NV center area depends on applying higher fluence and shorter pulse duration, regardless of the wavelength used. This trend can be experimentally observed in Fig. 2, which suggests that increasing the peak laser fluence leads to a greater likelihood of defect formation. To further support these results, however, it is essential to analyze the distribution of the observations. This relationship is illustrated in the correlation biplot presented in Fig. 6, which provides a simultaneous visualization of both variables and observations. Such representation allows for the joint interpretation of their orientations in the score and loading plots [33,54].

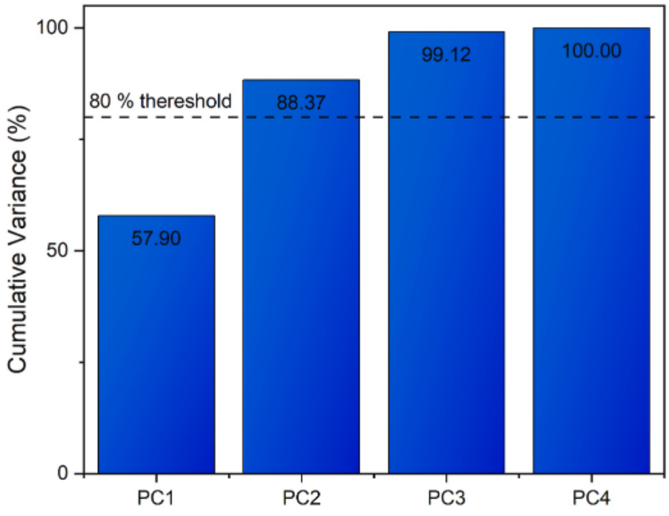


Fig. 4. Cumulative variance explained by the principal components (PCs). The first two PCs capture approximately 88.4 % of the total variance, surpassing the 80 % threshold. This indicates that the dataset's underlying structure can be effectively represented in two dimensions with minimal information loss.

Each point in the biplot corresponds to an observation from the original dataset. To help visualization of data clustering, a 95 % confidence ellipse was plotted. The data related to 1030, 515, and 343 nm are represented in red, green, and blue colors, respectively.

All the observations are primarily distributed along the PC1 axes, indicating effective clustering for the three excitation wavelengths, although 1030 nm shows the highest data dispersion. The overlap between the 343 and 515 nm groups suggests a significant similarity in their data distribution. The biplot also allows for the interpretation of the correlation between the original variables and the distribution of observation: data points oriented in the same direction as a given variable can be interpreted as exhibiting similar behavior.

Notably, the NV area is primarily associated with PC2, whereas PC1 most strongly describes all the other variables. These findings suggest that experimental conditions (increasing fluence while reducing the pulse duration) in the positive PC2 region yield larger NV areas. In contrast, the distribution along PC1 reveals that the data points corresponding to 343 nm and 515 nm lie in the negative region of this component, in opposition to the direction of the loadings associated with fluence and pulse duration. This suggests that optimal NV generation for these wavelengths is achieved at lower fluences and shorter pulse durations, supporting the experimental hypothesis that shorter wavelengths require lower energy inputs for efficient defect formation. Finally, the data points corresponding to 1030 nm are more dispersed and located along the positive axis of PC1, indicating a stronger dependence on fluence. This suggests that higher fluence levels are required to compensate for the lower photon energy at this wavelength, resulting in less efficient defect generation compared to shorter wavelengths.

These findings demonstrate the effectiveness of PCA as an exploratory tool for interpreting the dataset. By uncovering interdependencies among variables and identifying dominant trends, PCA enables informed decision-making for subsequent analyses, such as predictive modeling with ANNs [36,53,55].

3.2. Predictive measurements

A slight linear trend is observed in the predictions of the non-tuned model, as shown in Fig. 7(A) and (B). Prediction accuracy improves as the data points approach the identity line (y = x) and as R^2 approached unity. Likewise, MSE, MAE, and MAPE values closer to 0 indicate better performance. Table 2 shows that the external validation (Fig. 7(B)) yielded a slightly higher R^2 value than the internal validation (Fig. 7 (A)), with a difference of approximately 0.05. This low difference between training and testing is desirable, as it indicates good generalization and low overfitting. However, the MAPE of 35 % in the test set and 49 % in the train set using LOOCV reinforce that the non-tuned MLP was not capable of accurately predicting the formation of NV color centers and required further optimization.

Results for the tuned MLP are presented in Fig. 8(A) (internal validation) and Fig. 8(B) (external validation). The metrics summarized in Table 2 indicate that the tuned MLP model exhibits greater explanatory power compared to its non-tuned counterpart. Considering the R^2 of the test set, more than 90 % of the variation in experimental values can be explained by the predicted values of the tuned model. The difference between the validation scenarios is 0.06 for this R^2, indicating that the MLP does not suffer from significant overfitting and was capable of generalizing well the unseen data. A slight dispersion in the predictions for the external test set is expected, as this data was not involved in the training process.

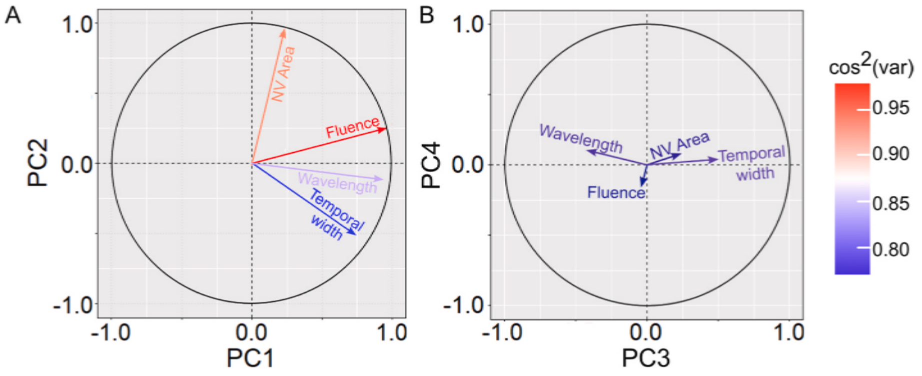


Fig. 5. Variable loadings for (A) PC1–PC2 and (B) PC3–PC4 planes. The projection highlights the contribution and directionality of each experimental parameter. Arrows closer to the unit circle indicate stronger correlations with the corresponding PCs. The PC3 and PC4 account for minor variance and reflect residual variation.

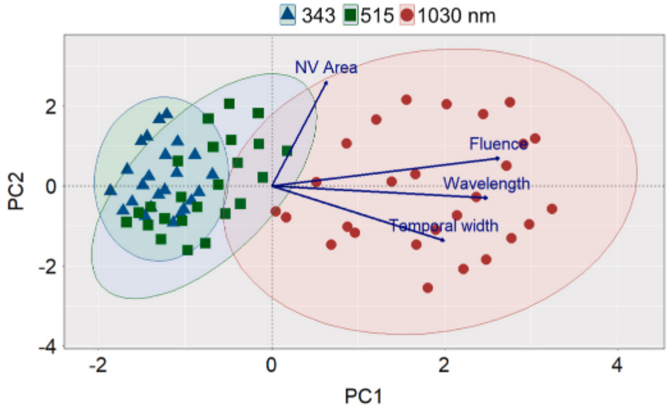


Fig. 6. Correlation biplot of the PCA scores and loadings, colored by laser wavelength. The PCA model reveals clear clustering of the data according to wavelength (343, 515, and 1030 nm), with partial overlap between 343 and 515 nm groups. The direction of the arrows reflects how each experimental parameter contributes to this separation.

Table 2 demonstrated that hyperparameter tuning significantly improved the performance of the MLP model, with improvements observed across all evaluation metrics. In external validation, the relative error reduction from the non-tuned to tuned model was 62 % for the MSE, 45 % for MAE and 42 % for MAPE. For internal validation, the reduction was 45 % for the MSE, 46 % for MAE and 65 % for MAPE. These results underscore the substantial impact of hyperparameter optimization on the predictive accuracy of the model, yielding more precise and reliable predictions. The optimized model was subsequently used to identify the combination of experimental parameters that maximizes or minimizes the percentage area of NV color center.

The ANN was supplemented with manually entered data at a fixed wavelength of 800 nm, while pulse duration and fluence were systematically varied within the bounds of their experimentally observed ranges, in steps of 50 fs and 0.05 J/cm^2, respectively. Table 3 shows that maximizing laser peak fluence and minimizing the excitation wavelength led to increased defect generation. However, pulse duration does not exhibit a clear influence, as it is predicted to be maximized in both scenarios, whether aiming to increase or decrease the percentage area of NV color centers. The discrepancy between PCA and ANN results regarding pulse duration likely stems from limitations in the dataset and methodological differences. The PCA emphasized global linear variance trends, while the ANN tried to capture nonlinear correlations that may require larger datasets to stabilize and produce accurate predictions.

Table 2
Performance metrics for non-tuned and tuned MLPRegressor.

Model	MLP			
	Default		Tuned	
Validation	Internal	External	Internal	External
MSE	0.17	0.16	0.09	0.06
MAE	0.31	0.33	0.17	0.18
MAPE	49 %	35 %	17 %	20 %
R [2]	0.76	0.81	0.87	0.93

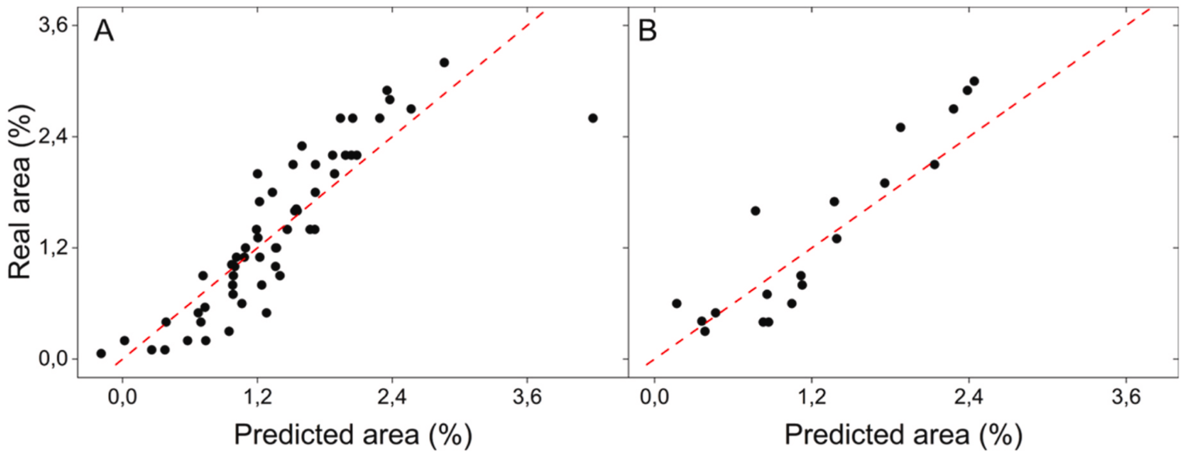


Fig. 7. Non-tuned prediction of the NV center generation using a MLPRegressor model. Comparison between the behavior of the real area (%) versus the predicted area (%) in the internal (A) and external (B) validation.

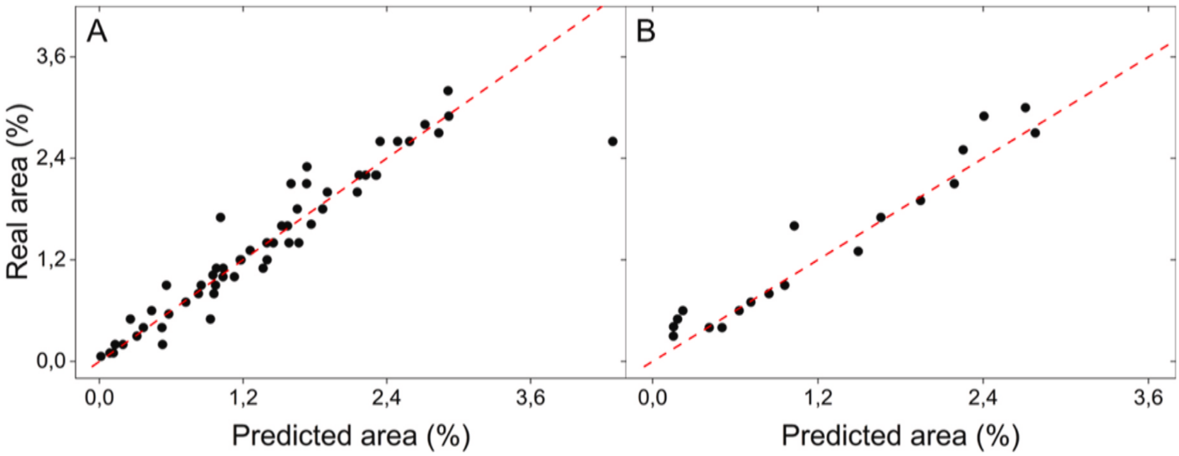


Fig. 8. Tuned prediction of the NV center generation using a MLPRegressor model. Comparison between the behavior of the real area (%) versus the predicted area (%) in the internal (A) and external (B) validation.

Table 3
Simulated optimal experimental values for maximizing and minimizing the NV color center generation via fs-laser micromachining.

Wavelength (nm)	Pulse duration (fs)	Fluence (J/cm^2)	NV area (%)
343	535	1.42	5.6663
515	535	1.42	4.8845
800	585	1.42	3.8287
1030	500	1.42	3.1534
343	1000	0.14	0.0967
515	1000	0.14	0.0009
800	235	0.19	0.0005
1030	985	0.51	0.0004

Therefore, the weak influence of pulse duration, as expressed in Tables 3 and is caused by the restricted data size and the complex interplay between fluence and temporal width in the proper NV center generation.

These results are consistent with the behavior observed in the original experimental data on vacancy generation (Fig. 2), as well as with the PCA-based interpretation. While previous analyses suggested an inverse relationship between pulse duration and defect formation, this effect appears to have a limited impact on the percentage area of NV center percent area when fluence is kept constant. Overall, the findings demonstrate that even a relatively simple ANN, when properly tuned, can extract relevant patterns from limited datasets and offer complementary insights into NV center generation experiments. However, it is important to note that the observed inverse correlation between excitation wavelength and NV center generation reflects dataset-specific trends rather than a universal physical law. Factors such as absorption depth, nonlinear ionization thresholds, and lattice damage mechanisms — none of which were explicitly modeled here — can significantly modulate this behavior, as well as a larger set of experimental data.

Despite higher laser fluences increasing the overall defect density, the presented results do not directly assess whether the generated defects preserve the quantum properties required for quantum photonics applications (e.g., spin lifetimes, linewidths). Further experimental studies, such as optically detected magnetic resonance (ODMR) measurements, would be valuable in determining whether NV centers produced with higher laser fluences exhibit the same quantum properties as those produced under other experimental conditions. This perspective constitutes an interesting direction for future work within this topic.

4. Conclusions

In summary, we applied computational tools to address a multivariate photonics problem involving interdependent experimental parameters. Dimensionality reduction and neural network methods were employed to analyze and predict nitrogen-vacancy (NV) center generation in diamond through femtosecond laser micromachining, with the goal of identifying key trends that could guide cost-effective optimization strategies. Principal component analysis, using a cumulative variance threshold of 80 %, reduced the data to a low-dimensional space in which two components explained 88 % of the total variance. The first principal component (PC1) alone accounted for 58 %, highlighting the dominant influence of fluence and the pulse duration over the excitation wavelength. The distribution of observations indicated that optimal NV center generation at 343 nm and 515 nm occurs at lower fluences, whereas 1030 nm requires significantly higher fluence levels. The implementation of a MLPRegressor (MLP) neural network with Leave-One-Out cross-validation (LOOCV) yielded a high coefficient of determination on unseen data ($R^2 = 0.93$) and low prediction errors for NV centers generations (MAPE = 20 %). The simulated predictions were consistent with the trends identified by PCA, demonstrating that the combined use of PCA and MLP offers a robust and complementary approach to data-driven analysis. We identified key experimental trends that can support the development of more efficient and cost-effective strategies for fs-laser generation of NV center in diamond.

CRediT authorship contribution statement

Murilo Neco Saraiva: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Orlando Marbello Ospina:** Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Lucas Konaka Nolasco:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft. **Renan Souza Cunha:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – review & editing. **Lucas Nunes Sales de Andrade:** Visualization. **Sergio Ricardo Muniz:** Visualization. **Cleber Renato Mendonca:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors are grateful to the São Paulo Research Foundation (FAPESP) (grants 2011/12399-0, 2015/20032-0, 2016/20886-1, 2018/11283-7, and 2020/08715-2), the National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of

Higher Education Personnel (CAPES, Finance Code 001), the Army Research Laboratory (grant W911NF-21-1-0362), and the Air Force Office of Scientific Research (grants FA9550-12-1-0028 and FA9550-23-1-0664).

Acknowledgments

We acknowledge financial support from the Sã o Paulo Research Foundation (FAPESP, grants 2011/12399-0, 2015/20032-0, 2016/20886-1, 2018/11283-7, and 2020/08715-2), the National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES, Finance Code 001), the Army Research Laboratory (grant W911NF-21-1-0362), and the Air Force Office of Scientific Research (grants FA9550-12-1-0028 and FA9550-23-1-0664). The authors also thank André L. Romero for technical assistance.

Data availability

Data will be made available on request.

References

[1] P.P. Shinde, S. Shah, A review of machine learning and deep learning applications, in: Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018, 2018, https://doi.org/10.1109/ICCUBEA.2018.8697857.

[2] S. Naeem, A. Ali, S. Anam, M.M. Ahmed, An unsupervised machine learning algorithms: comprehensive review, Inter. J. Computing Digital Sys. 13 (2023) 911–921.

[3] V. Nasteski, An overview of the supervised machine learning methods, HORIZONS. B 4 (2017) 51–62.

[4] M. Usama, et al., Unsupervised machine learning for networking: techniques, applications and research challenges, IEEE Access. 7 (2019) 65579–65615.

[5] J.E. van Engelen, H.H. Hoos, A survey on semi-supervised learning, Mach. Learn. 109 (2020) 373–440.

[6] A. Krogh, What are artificial neural networks? Nat. Biotechnol. 26 (2008) 195–197.

[7] R. Hecht-Nielsen, Theory of the backpropagation neural network, Neural Networks for Perception (1992) 65–93, https://doi.org/10.1016/B978-0-12-741252-8.50010-8.

[8] Popescu, M.-C. & Balas, V. E. Multilayer Perceptron and Neural Networks.

[9] J.E. Prilepsky, E. Manuylovich, P. Freire, S.K. Turitsyn, Artificial neural networks for photonic applications—from algorithms to implementation: tutorial, Adv. Opt Photon. 15 (3) (2023) 739–834, 739-834 15.

[10] Y. Xu, et al., Interfacing photonics with artificial intelligence: an innovative design strategy for photonic structures and devices based on artificial neural networks, Photon. Res. 9 (4) (2021) B135–B152. B135-B152 9.

[11] A.M. Hammond, R.M. Camacho, Designing integrated photonic devices using artificial neural networks, Opt. Express 27 (2019) 29620.

[12] R. Biswas, A.S. Kuar, S.K. Biswas, S. Mitra, Artificial neural network modelling of Nd:YAG laser microdrilling on titanium nitride—alumina composite, Proc Inst Mech Eng B J Eng Manuf 224 (2010) 473–482.

[13] J. Ciurana, G. Arias, T. Ozel, Neural network modeling and particle swarm optimization (PSO) of process parameters in pulsed laser micromachining of hardened AISI H13 steel, Mater. Manuf. Process. 24 (2009) 358–368.

[14] S.K. Dhara, A.S. Kuar, S. Mitra, An artificial neural network approach on parametric optimization of laser micro-machining of die-steel, Int. J. Adv. Manuf. Technol. 39 (2008) 39–46.

[15] A.A. Biswas, et al., Advancements in microneedle fabrication techniques: artificial intelligence assisted 3D-printing technology, Drug Deliv. Transl. Res. 14 (6) (2024) 1458–1479, 2024 14.

[16] A.D. Lantada, et al., Artificial intelligence aided design of microtextured surfaces: application to controlling wettability, Nanomater. 10 (2020) 2287. *2020 Page 2287* 10.

[17] M. Kianinia, I. Aharonovich, Diamond photonics is scaling up, Nat. Photonics 14 (2020) 599–600.

[18] C. Oncebay, J.M.P. Almeida, G.F.B. Almeida, S.R. Muniz, C.R. Mendonça, Localized Nitrogen-Vacancy centers generated by low-repetition rate fs-laser pulses, Diam. Relat. Mater. 130 (2022).

[19] J.R. Maze, et al., Properties of nitrogen-vacancy centers in diamond: the group theoretic approach, New J. Phys. 13 (2011) 025025.

[20] V.M. Acosta, A. Jarmola, E. Bauch, D. Budker, Optical properties of the nitrogen-vacancy singlet levels in diamond, Phys. Rev. B Condens. Matter. 82 (2010) 201202.

[21] F. Jelezko, J. Wrachtrup, Single defect centres in diamond: a review, Physica Status Solidi (A) Appl. Mater. Sci. 203 (2006) 3207–3225.

[22] J. Wrachtrup, F. Jelezko, Quantum information processing in diamond. https://arxiv.org/pdf/quant-ph/0510152, 2005.

[23] Y. Doi, et al., Pure negatively charged state of the NV center in <span class="aps-inline-formula">*n*</span>-type diamond, Phys. Rev. B 93 (2016) 081203.

[24] Y.C. Chen, et al., Laser writing of scalable single color centers in silicon carbide, Nano Lett. 19 (2019) 2377–2383.

[25] Y.-C. Chen, et al., Laser writing of individual nitrogen-vacancy defects in diamond with near-unity yield, Optica 6 (2019) 662.

[26] G. Balasubramanian, A. Lazariev, S.R. Arumugam, D. wen Duan, Nitrogen-vacancy color center in diamond-emerging nanoscale applications in bioimaging and biosensing, Curr. Opin. Chem. Biol. 20 (2014) 69–77.

[27] M.W. Doherty, et al., The nitrogen-vacancy colour centre in diamond, Phys. Rep. 528 (2013) 1–45.

[28] V.V. Kononenko, et al., Nitrogen-vacancy defects in diamond produced by femtosecond laser nanoablation technique, Appl. Phys. Lett. 111 (2017).

[29] C. Oncebay, J.M.P. Almeida, G.F.B. Almeida, S.R. Muniz, C.R. Mendonca, Localized nitrogen-vacancy centers generated by low-repetition rate fs-laser pulses, Diam. Relat. Mater. 130 (2022).

[30] E.V. Levine, et al., Principles and techniques of the quantum diamond microscope, Nanophotonics 8 (2019) 1945–1973.

[31] M.E. Trusheim, D. Englund, Wide-field strain imaging with preferentially aligned nitrogen-vacancy centers in polycrystalline diamond, New J. Phys. 18 (2016) 123023.

[32] L.K. Nolasco, L.N.S. de Andrade, S. Pratavieira, S.R. Muniz, C.R. Mendonça, Optimization of Nitrogen-vacancy center production using ultrashort laser pulses, Appl. Surf. Sci. 713 (2025) 164318.

[33] F.L. Gewers, et al., Principal component analysis: a natural approach to data exploration, ACM Comput. Surv. 54 (2021).

[34] M. Ringnér, What is principal component analysis? Nat. Biotechnol. 26 (2008) 303–304.

[35] Bharadiya, J. P. A tutorial on principal component analysis for dimensionality reduction in machine learning. https://doi.org/10.5281/ZENODO.8002436 doi: 10.5281/ZENODO.8002436.

[36] R.D.O. Santos, et al., Principal component analysis and factor analysis: differences and similarities in nutritional epidemiology application, Rev. Bras. Epidemiol 22 (2019) e190041.

[37] N. Salem, S. Hussein, Data dimensional reduction and principal components analysis, Procedia Comput. Sci. 163 (2019) 292–299.

[38] R. Silva, P. Melo-Pinto, A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries, Appl. Soft Comput. 113 (2021) 107889.

[39] K. Kumar, G.S.M. Thakur, Advanced applications of neural networks and artificial intelligence: a review, Int. J. Inf. Technol. Comput. Sci. 4 (2012) 57–68.

[40] E. Kim, et al., Innovative strategies for protein content determination in dried laver (Porphyra spp.): evaluation of preprocessing methods and machine learning algorithms through short-wave infrared imaging, Food Chem. X 23 (2024) 101763.

[41] M.M. Ahsan, M.A.P. Mahmud, P.K. Saha, K.D. Gupta, Z. Siddique, Effect of data scaling methods on machine learning algorithms and model performance, Technologies 9 (2021) 52. *2021 Page 52* 9.

[42] M. Aswin Krishna, A. Kedawat, A. Bansal, R. Suresh, Feature engineering for neural network-based oscillation detection in process industries, Computer Aided Chem. Eng. 51 (2022) 1153–1158.

[43] R. C R, P. C, S. D, Evaluating Deep learning with different feature scaling techniques for EEG-based music entrainment brain computer Interface, e-Prime - Advances in Electrical Engineering, Electronics and Energy 7 (2024) 100448.

[44] T.D.K. Thara, P.S. Prema, F. Xiong, Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques, Pattern Recognit. Lett. 128 (2019) 544–550.

[45] E. Elgeldawi, A. Sayed, A.R. Galal, A.M. Zaki, Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis, Informatics 2021 8 (2021) 79, 79 8,.

[46] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (2020) 295–316.

[47] R. Hossain, D. Timmer, Machine Learning Model Optimization with Hyper Parameter Tuning Approach, 2021.

[48] T.T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recogn. 48 (2015) 2839–2846.

[49] G.C. Cawley, N.L.C. Talbot, Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers, Pattern Recogn. 36 (2003) 2585–2592.

[50] A new typology design of performance metrics to measure errors in machine learning regression algorithms, Interdiscipl. J. Inf. Knowl. Manag. 14 (2019) 45–76.

[51] A. De Myttenaere, B. Golden, B. Le Grand, F. Rossi, Mean Absolute percentage error for regression models, Neurocomputing 192 (2016) 38–48.

[52] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831.

[53] E.C. Malthouse, Limitations of nonlinear PCA as performed with generic neural networks, IEEE Trans. Neural Network. 9 (1998) 165–173.

[54] H. Shafizadeh-Moghadam, Fully component selection: an efficient combination of feature selection and principal component analysis to increase model performance, Expert Syst. Appl. 186 (2021) 115678.

[55] C. Fan, N. Zhang, B. Jiang, W.V. Liu, Using deep neural networks coupled with principal component analysis for ore production forecasting at open-pit mines, J. Rock Mech. Geotech. Eng. 16 (2024) 727–740.