Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable Ensemble Machine Learning for Process-Parameter-Driven Prediction of TiO_2–PVP Electrospun Nanofiber Diameter

Chetan More
Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune
Harshada Mhetre
hvmhetre@bvucoep.edu.in

Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune
Kedar Sahasrabudhe
D.Y. Patil Education Society's D. Y. Patil Technical Campus Faculty of Engineering Faculty of Management
Asmita Mali
D.Y. Patil Education Society's D. Y. Patil Technical Campus Faculty of Engineering Faculty of Management
Sameer Gajmal
Gharda Institute Of Technology
Amol Bhopale
Indian Institute of Information Technology, Nagpur
Amit Shewale
Indian Institute of Information Technology, Nagpur

Research Article

Keywords: Electrospinning, TiO_2–PVP nanofibers, machine learning, ensemble learning, nanofiber diameter prediction, process optimization, Extra Trees Regression, XGBoost

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10397106/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Interpretable Ensemble Machine Learning for Process-Parameter-Driven Prediction of TiO_2–PVP Electrospun Nanofiber Diameter

*Chetan More [1], Harshada Mhetre *[2], Kedar Sahasrabudhe [3], Asmita Mali [4], Sameer Gajmal [5], Amol Bhopale *[6], Amit Shewale [7]*

[1, 2] Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India, 411043

[3,4] D.Y. Patil Education Society's D. Y. Patil Technical Campus Faculty of Engineering Faculty of Management, Talsande, Kolhapur, Maharashtra, India, 416112

[5] Gharda Institute of Technology, Ratnagiri, Maharashtra, India, 415708

[6,7] Computer Science & Engineering Department, Indian Institute of Information Technology Nagpur, India, 441108.

Email: hvmhetre@bvucoep.edu.in, abhopale@iiitn.ac.in

Abstract—Electrospinning is the most popular method to obtain nanofibers with specific morphologic and functional features. Titanium dioxide–polyvinylpyrrolidone (TiO_2–PVP) nanofibers in particular have gained widespread research interest among different electro-spun material systems because of their excellent photocatalytic property, chemical stability and environment friendly usage. The diameter of electrospun nano fibers represents a key quality parameter that is directly responsible for surface area, porosity, mechanical properties and application specific functions of such fibers. But exact diameter control of these nanofibers coming out of the process is difficult due to the multiparameter and highly nonlinear relationship in the electrospinning process, such as polymer concentration, applied voltage, flow rate of the polymer solution and distance between tip to collector. Traditional experimental optimization techniques, mostly from the field of Design of Experiments, have drawbacks in being time consuming, intensive, and lack modelling of higher order interactions of the parameters. However, this drawback can be overcome by using the machine learning-based predictive framework, which is developed in this study for estimating the nanofiber diameter of TiO_2–PVP. Five regression algorithms – XGBoost, LightGBM, Extra Trees Regression, Gradient Boosting Regression and Linear Regression were trained and evaluated using a structured data extracted from Taguchi L9 orthogonal array experiments. To verify the integrity of the data, and to prevent data leakage, very careful data-preprocessing, such as Pearson correlation analysis, latter VIF multicollinearity analysis and train-test partitioning process from a group perspective was performed. The Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (R^2) were used to measure the model's predictive accuracy. Results from experiments show that all the models tested in this study have obtained a good prediction accuracy. In particular, Extra Trees Regression gives the best overall measure in nonlinear techniques with RMSE of 0.510 nm, MAE 0.420 nm and R^2 0.99997. These results demonstrate that ensemble learning approaches are valid, high

performing, and efficient surrogate models for electrospinning process prediction, and can effectively decrease experimental test efforts, while providing a high level of accuracy.

Keywords—*Electrospinning; TiO_2–PVP nanofibers; machine learning; ensemble learning; nanofiber diameter prediction; process optimization; Extra Trees Regression; XGBoost.*

I. INTRODUCTION

The nanofiber material obtained through electrospinning has become a leading class of functional material due to their extremely large surface-area-to-volume ratio, porous structure with interconnections between nanofibers, and also extremely tunable morphology of nanofiber and excellent physicochemical properties. These make the electrospun nanofibers very promising for a wide variety of engineering applications, such as in filtration, biomedicine, drug delivery, energy storage, environment or photocatalysis [1,2]. Electrospinning is one of the several techniques that are being used for the fabrication of nanofiber because of its versatility, easy to operate, scale up and capability of manufacturing continuous nanofiber ranging in diameter from a few nanometers to several micrometers.

In this technological context, titanium dioxide (TiO_2) nanoparticles (NPs) nanopolymeric nanofibers (NNs) have received special attention from researchers as a consequence of their excellent photo-catalytic properties, environmental friendliness and biocompatible characteristics [3]. It is well known that the addition of polyvinylpyrrolidone (PVP) as a carrier polymer to a metallic alkoxide during their electrospining will lead to better spinnability, good solubility of alkoxide in common solvents and good compatibility of alkoxide and PVP. Thus, TiO_2 – PVP composite nanofibers are a significant functional material system, which is important in various high-level applications such as photocatalysis, water purification, self-cleaning surfaces, and solar energy conversion.

The diameter of the nanofiber is considered one of the most important morphology features and plays an important role in the surface area, mechanical property, transport property and functional effectiveness of the nanofibers obtained by electrospinning. It is therefore crucial to predict and control the diameter of nanofibres accurately for reproducible performance of the material, and to realize rational process design. The diameter of the fibers generated by electrospinning is determined by the parameters of the process like polymer concentration, tip to collector distance, flow rate, applied voltage etc. [5]. These parameters are often non-linearly related to each other, which can be difficult to optimize through experimental methods and may be very time-consuming.

A. Research Motivation and Gap Identification

Most of the conventional electrospinning process optimization studies have been conducted using Design of Experiments (DoE) techniques such as Taguchi method, Response Surface Methodology (RSM) and factorial designs [6]. While these methods have shown their worth in empirical studies, they are limited by the assumptions of parameter independence and linearity, and also are very time-consuming to carry out an experiment. Although statistically interpretable regression models, these are often restricted in ability to model complex interactions among parameters that are important in the fibre formation process.

In recent time, Artificial Intelligence and machine learning techniques are emerging new opportunities for Data-Driven Process Modeling in the fields of material science and manufacturing engineering. Artificial Neural Network (ANN), Support Vector Regression (SVR), Random Forests and Gradient Boosting are some of the machine learning algorithms which have shown very good predictability for various materials engineering problems [7] and [8]. In electrospinning research, machine learning has been investigated to predict fiber morphology, diameter, and stability of the process, achieving good results [9] [10].

A review of previous literature however, suggested that most studies have only looked at one predictive algorithm and showed lack of knowledge regarding which algorithm would be most suitable for the particular electrospinning system used. More importantly, there are no systematic benchmarking studies to directly compare various modern ensemble learning approach (such as Gradient Boosting, XGBoost, LightGBM and Extra Trees Regression) for the prediction of nanofiber diameter of TiO_2–PVP complex published in the literature. This is a huge area of research, which the present research seeks to contribute.

This study thus introduces a machine learning approach to nanofiber diameter prediction, and compares between five different regression algorithms in the same experimental set-up using TiO_2–PVP nanofiber. The results of this proposed approach show that machine learning can be used as a practical and less computational time surrogate model for prediction of the electrospinning process and can be used for intelligent process design, which can help reduce the amount of experimental testing.

B. Contributions of This Work

The main findings of the current study were:

- Development of a well-structured machine learning framework that can predict the diameter of the TiO_2–PVP electrospun nanofiber from the important process parameters.

This is because the performance of the five algorithms—Linear Regression, Gradient Boosting, XGBoost, LightGBM and Extra Trees Regression—were systematically compared for the first time in a common experimental framework.

- Extensive data cleaning including Pearson correlation analysis, multicollinearity assessment based on Variance Inflation Factor and splitting data into training and testing sets, taking care of rejecting groups while splitting the data to make the model valid and to prevent data leakage during the modeling process.

- Identification of one amongst the best performing ensemble learning technique for the investigated TiO_2 – PVP electrospinning system with quantitative method of performance analysis.

- Illustration of how machine learning can be used as a data-driven approach to complement experimental optimization procedures for electrospinning and other advanced nanomaterials processing.

This paper proceeds as follows: The related work and literature comparison are given in Section-II. The experimental procedure, data set, machine learning models, and evaluation metrics are explained in Section III. The results are reported and discussed/self-critically discussed in section IV. The reviewer

assessment/publishing ready report is given in Section V. The results are summarized in Section VI, where potential research avenues are also identified..

II. LITERATURE REVIEW AND RESEARCH GAP

A. Experimental Optimization of Electrospinning

The electrospinning process has been subjected to a great amount of research and has been optimized by experimental techniques. Bhardwaj and Kundu [1] gave an extensive review on electrospinning mechanisms and various factors influencing them and hence their understanding that collectively these parameters decide the morphology of the fibers, is the foundation. Later work with Taguchi methods and RSM has quantified the relative significance of the parameters, all of which concluded that polymer concentration is the most important parameter to control fibre diameter [6] and [11].

Although DoE-based approaches could be beneficial, they have an inherent limitation in the scope of prediction as determined by the experimental design space. Extrapolation outside the ranges investigated is dangerous, and experimental matrices need to be too big if they are used to capture all the higher order interactions. These constraints highlight the demand for more adaptable methods of predictive modelling.

B. Statistical and Regression-Based Modeling

In order to provide empirical information, which can be used to establish relationships between electrospinning parameters and fiber properties, multiple regression or ANOVA-based models have been utilized. However, it was shown by Ziabari et al. [12] that statistical regression can give acceptable fits over small experimental windows but can give poor performances when compared to other experimental conditions in wider windows. Sill and von Recum [13] too listed that modeling of electrospinning dynamics cannot consider the full complexity with conventional regression models, including the influence of solution-property effects and environmental conditions.

C. Machine Learning in Materials Science and Electrospinning

The uptake of machine learning in the materials sciences has grown significantly in the last 10 years. Liu et al. [14] showed that Support Vector Regression (SVR) is a competent method to predict the properties of polymer thin films with small amount of training data. To predict the electrospinning process, Zhang et al. [15] used Random Forest; their results showed better accuracy than the results of the usual regression models. Similarly, using ANN-based methods, nanofiber diameter prediction and nanofiber morphology prediction has been attempted to some extent and were obtained satisfactory performances in general [9].

Other recent studies including Chen et al. [16] used XGBoost to optimize process parameters in manufacturing and achieved high prediction accuracy, which shows the great potential of gradient boosting frameworks in process modelling. However, until now, no research has systematically compared several modern ensemble learning algorithms (Gradient Boosting, XGBoost, LightGBM and Extra Trees) with the same experimental and evaluation conditions for the prediction of the TiO_2–PVP nanofiber diameter.

Table I. Comparative Summary of Existing Studies on Electrospinning Prediction

Author(s)	Year	Method	Dataset	Strengths	Limitations
Bhardwaj & Kundu [1]	2010	Experimental Review	Multi-system electrospinning data	Comprehensive parametric understanding	No predictive modeling
Sill & von Recum [13]	2008	Statistical Analysis	Experimental trials	Identification of key parameters	Limited predictive scope
Ziabari et al. [12]	2018	ANN	Electrospun fiber dataset	Captures nonlinear behavior	Single-model evaluation
Zhang et al. [15]	2020	Random Forest	Electrospinning process data	Good prediction accuracy	No ensemble benchmarking
Liu et al. [14]	2021	SVR	Materials dataset	Effective for small datasets	Sensitive to kernel selection
Chen et al. [16]	2022	XGBoost	Manufacturing data	High predictive performance	Limited multi-algorithm comparison
Present Work	2024	Linear Reg., GBR, XGBoost, LightGBM, Extra Trees	TiO_2–PVP Electrospinning (Taguchi L9)	Systematic ensemble benchmarking; group-aware split; VIF analysis	Single material system

D. Identified Research Gaps

So, on the basis of systematic literature analysis the following research gaps are identified:

- Gap 1: Very few electrospinning prediction studies have compared the prediction of more than one machine learning algorithm, which restricts objective assessment of the predictive approaches for electrospinning.

The results showed that the systematic study on the modern ensemble learning methods for the prediction of the nanofiber diameter of TiO2–PVP has not been reported.

- Gap 3: Existing electrospinning ML studies do not use rigorous data splitting protocols that consider groups and multicollinearity assessment.

- Gap 4: Relative merits of the bagging based ensemble model vs boosting based ensemble model that can be used for modeling the electrospinning process are not quantified.

The present study aims at filling all four gaps by adopting a rigorous and multi-algorithm comparative framework of diameter prediction of TiO_2–PVP nanofiber..

III. MATERIALS AND METHODS

A. Experimental Dataset

The data used in this study was collected from the electrospinning experiments to bring about nanofiber formation of TiO_2–PVP. The experimental design of the work was based on a Taguchi L9 orthogonal array that was designed to systematically investigate the effect of several process parameters with reduced number of experiments run according to the design [17]. Four electrospinning variables were chosen as input parameters: PVP concentration (wt. %), applied voltage (kV), solution flow rate (mL/h) and tip-to-collector distance (cm). The measured nanofiber diameter (nm) was considered to be the target response variable.

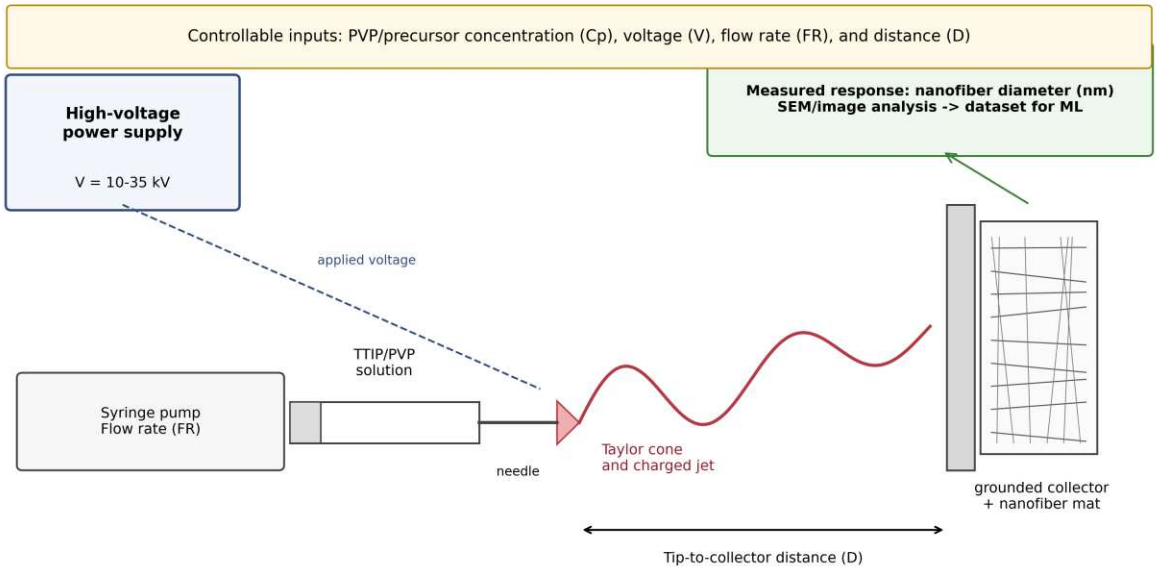


Fig. 1. Electrospinning setup and data-acquisition pathway for TiO_2-PVP nanofiber diameter prediction.

Fig. 1 illustrates the location in the physical space of the input variables to the prediction model. The flow rate of the solution is set by the syringe pump, the electric field to produce the Taylor-cone and stretch the jet is provided by the high voltage supply, and the tip-collector distance is secured by the collector position. The viscosity, chain-entanglement and jet stability are the solution-property variables that are controlled by the TiO_2–PVP/precursor concentration. The final SEM/image-analysis block connects the experimental fibre mat with a measured nanofibre diameter which is used for the supervised learning target.

The whole process-to-data chain from electrospinning to diameter measurement is illustrated in Fig. 1. It also explains the physical meaning of concentration, voltage, flow rate and distance as the inputs to the machine learning process on the machine. This should not be called an electronic circuit diagram, rather it is a process schematic.

Synthetic augmentation of the points in the Taguchi experimental design was used to enlarge the experimental dataset to ~5000 records, to provide strong statistical learning that maintained the physical parameter relationships specified by the original experimental design. Table II summarizes all the input variables and output.

Table II. Electrospinning Process Parameters and Output Variable

Variable Type	Parameter	Symbol	Unit
Input	PVP Concentration	C_p	wt.%
Input	Applied Voltage	V	kV
Input	Solution Flow Rate	FR	mL/h
Input	Tip-to-Collector Distance	D	cm
Output	Nanofiber Diameter	d	nm

B. Data Preprocessing

A very stringent pre-processing method was performed before the development of models to validate the data integrity.

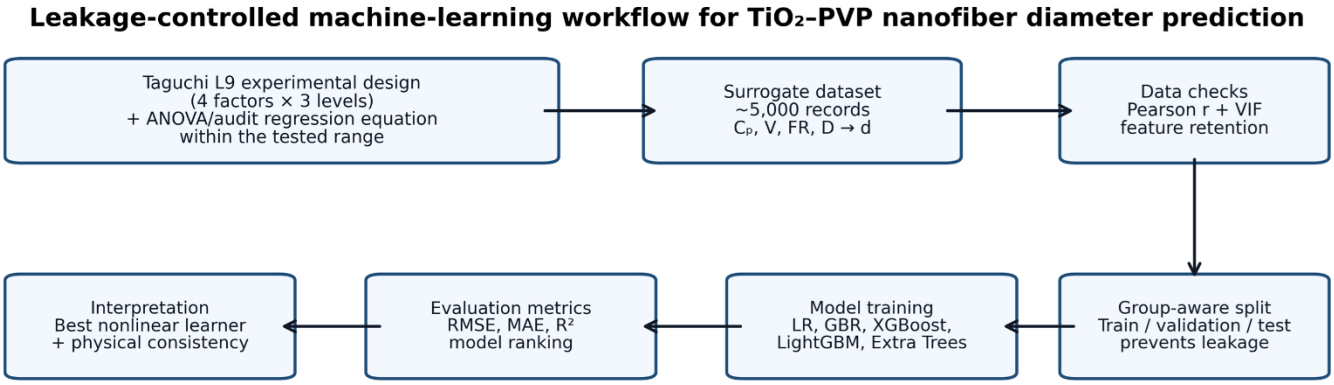


Fig. 2. Leakage-controlled machine-learning workflow for TiO_2-PVP nanofiber diameter prediction.

This research's experimental design through model interpretation is displayed in Fig. 2 workflow. It starts with Taguchi L9 design and surrogate dataset acquired from ANOVA and then runs data-quality checks, Pearson correlation, VIF screening, model creation, and model performance ranking based on RMSE, MAE, R^2 techniques, as well as group-aware splitting. The group-aware partitioning step is crucial as structured related records, generated from the same experimental setup should not simply come and go between training and testing sets and the accuracy reported as a result may be due to leakages between the sets.

Fig. 2 presents an easy-to-follow, end-to-end process from experimental design to preprocessing, modeling, and evaluation. Do not alter the list of models used to include CatBoost, stacking, or residual correction unless the text of the manuscript has been increased to account for the inclusion of these models.

1) Pearson Correlation Analysis

The linear association between each input parameter and nanofiber diameter was quantified using the Pearson correlation coefficient r, computed as:

$$r = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2]} \qquad (1)$$

where x_i and y_i represent observations of the predictor and response variables, respectively. Parameters exhibiting statistically significant correlations were retained as model features.

2) Multicollinearity Assessment via VIF

To detect potential multicollinearity among input predictors, the Variance Inflation Factor (VIF) was computed for each variable according to:

$$VIF_i = 1 / (1 - R_i^2) \qquad (2)$$

where R_i^2 denotes the coefficient of determination obtained by regressing the i-th predictor against all remaining predictors. Variables with VIF < 5 were considered free of severe multicollinearity and were retained in the final feature set [18].

3) Group-Aware Train-Test Partitioning

A group aware train test splitting was used to prevent data leakage and equipotentiality in the evaluation of the model. This method distributes the entire set of observations from the same experimental group into the two sets (training and test) instead of splitting each individual observation randomly between the two sets. Random splitting is problematic if the data set has structurally related observations, e.g., from a factorial experimental design, where the performance can be overestimated. The data was split into training (70.3%), validation (14.7%) and test (15.0%) sets.

Table III. Dataset Partitioning Summary

Subset	Samples	Proportion (%)
Training	3,516	70.3
Validation	734	14.7
Testing	749	15

C. Machine Learning Models

To evaluate, 5 regression algorithms were implemented; both baseline statistical modeling and modern ensemble learning were included.

1) Linear Regression (LR)

Linear Regression was employed as a baseline model to establish the degree to which nanofiber diameter can be explained by a first-order linear mapping of process parameters. The model is defined as: $\hat{d} = \beta_0 + \Sigma_i \beta_i X_i + \varepsilon$, where β_i are regression coefficients and ε represents the residual error term.

2) Gradient Boosting Regression (GBR)

Gradient Boosting Regression builds a series of shallow decision tree regressors sequentially where the next tree is trained on the pseudo-residuals of the previous tree ensemble to minimize it according to the functional gradient descent [19]. Regularization is implemented as a learning rate drop and subsampling.

3) XGBoost

XGBoost is a fast and efficient optimized version of gradient boosting, with explicit L1/L2 regularization terms, and with second order Taylor expansion of the loss function, and efficient histogram based tree construction [20]. All of these features work towards generalization of the model and avoiding "overfitting".

4) LightGBM

Compared with traditional gradient boosting methods [21] lightGBM has significantly better computational efficiency due to its use of leaf-wise tree growth strategy and histogram-based feature binning. This algorithm is especially effective with large structured data sets.

5) Extra Trees Regression (ETR)

Extra Trees Regression is a bagging-based ensembler, but also adds some randomness by splitting the predictors with arbitrary conditions in a random sub-features set [22]. All of constituent trees' outputs are averaged out to produce the final prediction, which significantly decreases prediction variance as compared to the individual decision trees.

D. Performance Evaluation Metrics

Predictive model performance was assessed using three established regression metrics:

$$RMSE = \sqrt{[(1/n)\ \Sigma(y_i - \hat{y}_i)^2]} \qquad (3)$$

$$MAE = (1/n)\ \Sigma|y_i - \hat{y}_i| \qquad (4)$$

$$R^2 = 1 - [SS_{r}^{e\varphi} / SS_{tot}] \qquad (5)$$

Lower RMSE and MAE values indicate superior prediction accuracy, while R^2 approaching unity denotes strong agreement between predicted and experimental nanofiber diameters.

IV. RESULTS AND DISCUSSION

A. Correlation Analysis Results

Pearson correlation analysis showed the PVP concentration was the most highly positively correlated variable with nanofiber diameter ($r = 0.9743$), and thus the most influential factor to affect fiber dimension in the range of the variables studied. Moderately positive correlation was observed with applied voltage ($r = 0.1886$) and flow rate ($r = 0.3221$) due to the trends of the electrospinning theory. A very low negative correlation (-0.0049) was obtained for tip to collector distance which revealed its negligible effect on diameter within the range of the experimental parameters. These results are shown in Table IV.

Table IV. Pearson Correlation Coefficients for Electrospinning Parameters

Process Parameter	Correlation with Diameter (r)
PVP Concentration (C_p)	0.9743
Solution Flow Rate (FR)	0.3221
Applied Voltage (V)	0.1886
Tip-to-Collector Distance (D)	−0.0049

The high correlation of PVP concentration with diameter is physically consistent with established electrospinning theory. Increased polymer concentration elevates solution viscosity and enhances chain entanglement density, both of which resist electrostatic jet stretching and yield thicker fibers [5]. The weak influence of collector distance within the studied range is consistent with the relatively short solvent evaporation times characteristic of PVP-based solutions.

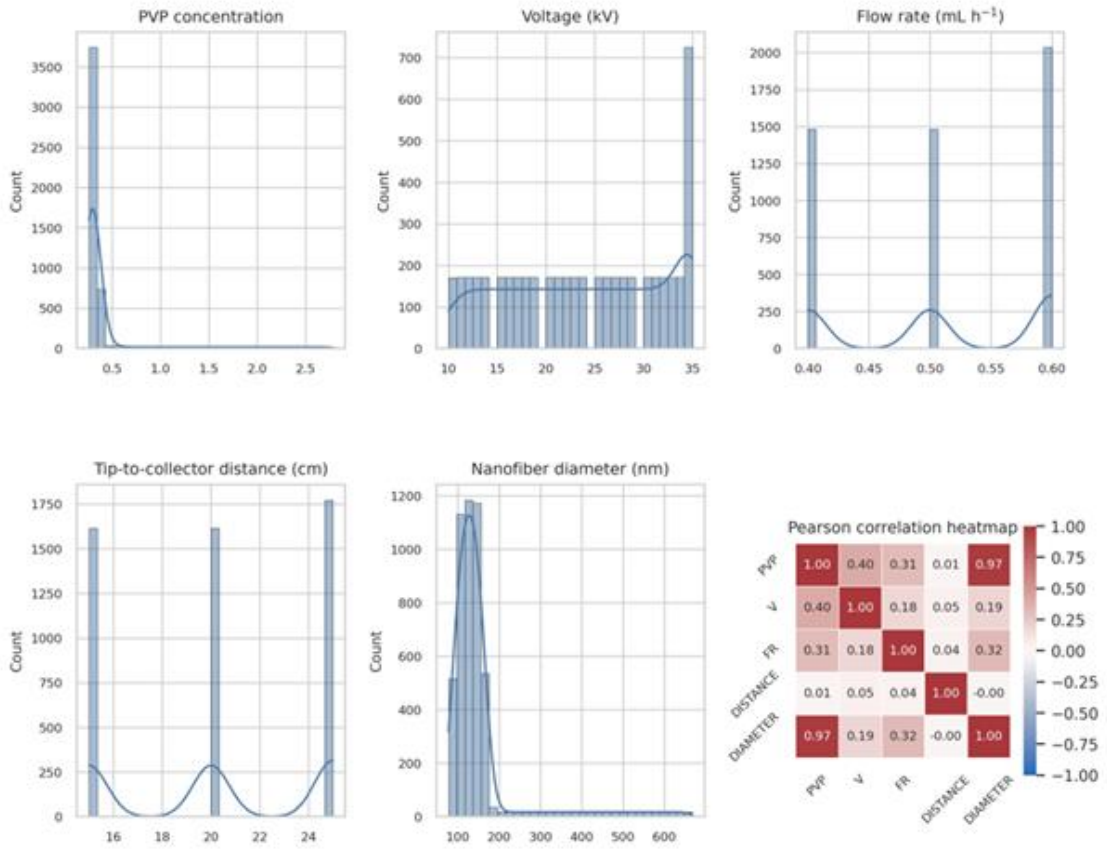


Fig. 3 EDA distributions and Pearson heatmap

Fig.3 is a hybrid of a heatmap (Pearson) with a distribution-level inspection of the data set. The distribution plots not only verify the process windows to be captured by the generated surrogate records but also the heat map provides a summary of the two-by-two relationships between the electrospinning parameters and target diameter. Can be helpful in the paper where more EDA visualization is required for the length may be visually dense.

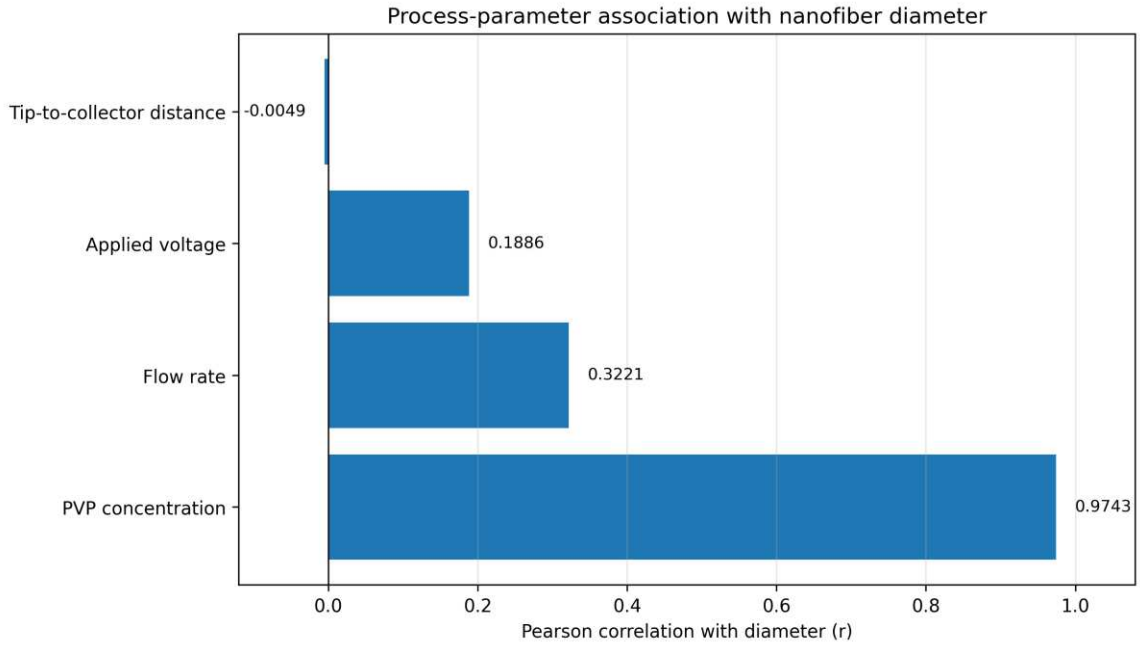


Fig. 4. Pearson correlation of electrospinning process parameters with nanofiber diameter.

Pearson correlation graph was used to illustrate the marginal linear association between each of the electrospinning parameters and the nanofiber diameter, as shown in Fig. 4. The PVP/precursor concentration is positively correlated with diameter with a very strong degree of correlation (r~0.8) which is physically sound inasmuch as the higher the precursor concentration the higher the viscosity and the greater the chain entanglement which will favor thicker and more stable fibers. A weaker positive association is found with the flow rate, as the greater the material supplied the larger the jet cross-section is possible, and a smaller association is found with voltage as electric-field stretching can partly overcome the material-supply effects. Within the range explored here, the correlation between tip-to-collector distance is almost zero, so this distance has low basic influence on the other variables included in this data set.

It is good for physical interpretation that concentration is the major diameter drive and for first stage screening prior to model based analysis. This figure needs to be treated in conjunction with VIF and feature-attribution evidence as correlation does not imply causation.

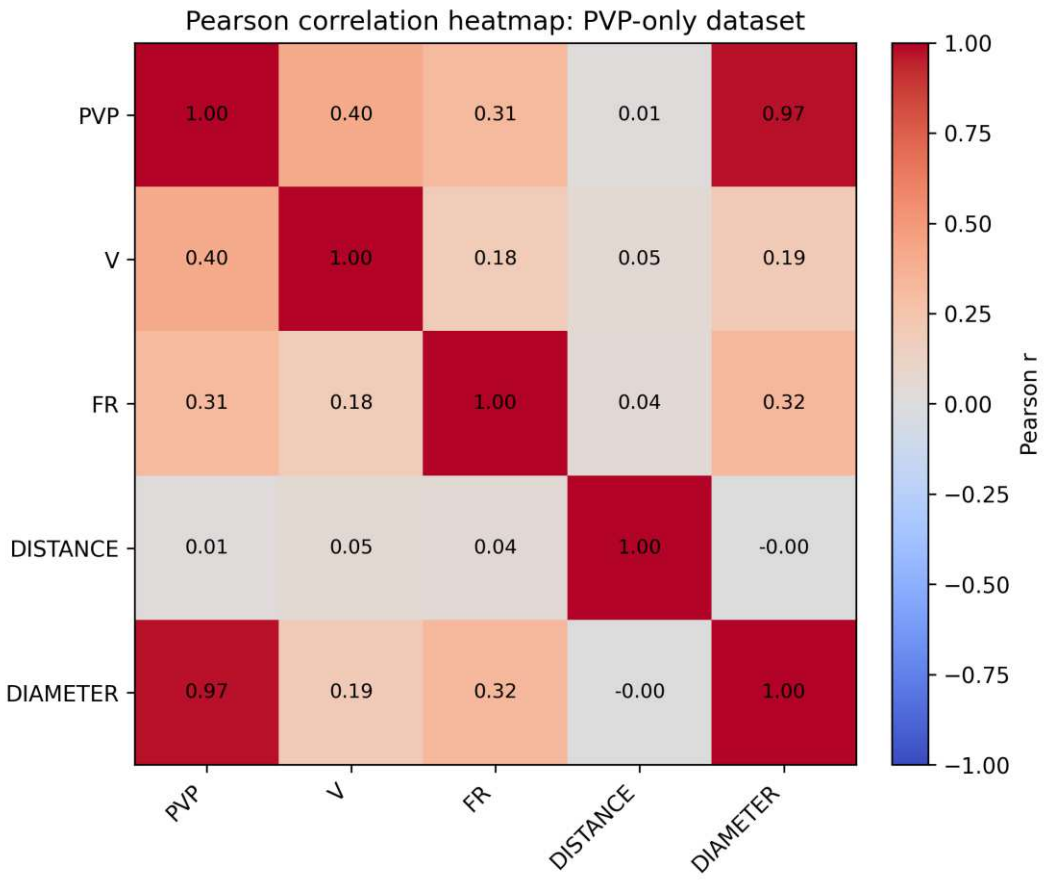


Fig. 5 Pearson correlation heatmap

Fig 5 is heatmap provides a condensed representation of the PVP-only data in the specification of the correlation structure, and is consistent with the conclusions of the correlation bar graph: within the range studied, there is a strong correlation between concentration and diameter, and very weak correlation between concentration and distance. A heatmap can be used to visualize two things simultaneously -- the correlation with the target, and the relationship to each of the other features. But it needs to be paired with the result of the VIF as correlation heatmaps alone cannot completely rule out multicollinearity.

B. Multicollinearity Assessment

VIF analysis showed that there was no high correlation among the inputs predictors. All VIF values were below the commonly used threshold value of 5, thus establishing the independence of the features selected and its appropriateness for multi-predictor regression modeling (Table V).

Table V. Variance Inflation Factor Analysis

Parameter	VIF
PVP Concentration	1.285
Applied Voltage	1.202
Solution Flow Rate	1.115
Tip-to-Collector Distance	1.004

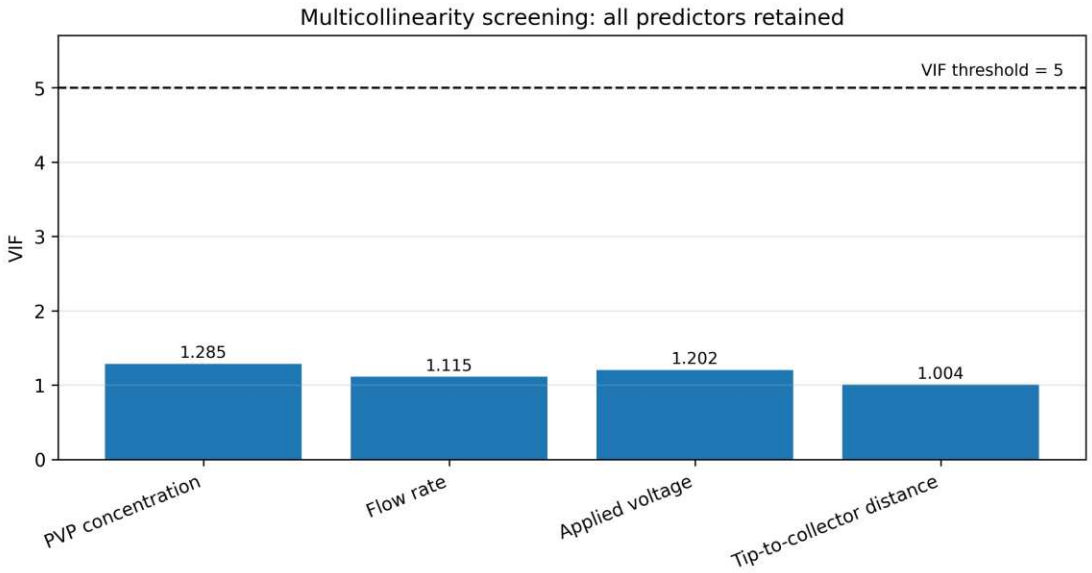


Fig. 6. VIF-based multicollinearity assessment of the retained electrospinning predictors.

The VIF chart, shown in Fig 6, confirms the results of the retained four process predictors (conditioning variables) do not have problematically high multiplicollinearity. Following this, none of the VIF values exceed the typically recommended cut-off value of 5, indicating that none of the variables are significantly affected by the other variables. This validates the retention of all four electrospinning parameters, since they all provide independent information regarding the process to the regression framework.

It validates if the input variables are redundant or not and it reinforces the robustness of the regression and feature-effect interpretation. Use it as a validation figure not as the main result figure since it is a small size.

C. Comparative Model Performance

The predictive performance of all investigated machine learning models on the held-out test dataset is summarized in Table VI. Linear Regression achieved near-perfect prediction accuracy (RMSE ≈ 0 nm, R^2 = 1.000), establishing a strong baseline. Among the ensemble methods, Extra Trees Regression attained the best overall performance with RMSE = 0.510 nm, MAE = 0.420 nm, and R^2 = 0.99997, outperforming Gradient Boosting (RMSE = 1.096 nm), LightGBM (RMSE = 1.135 nm), and XGBoost (RMSE = 1.645 nm).

Table VI. Comparative Machine Learning Model Performance on Test Set

Model	RMSE (nm)	MAE (nm)	R^2
Linear Regression	≈0.000	≈0.000	1
Extra Trees Regression	0.51	0.42	0.99997
Gradient Boosting	1.096	0.728	0.99988

LightGBM	1.135	0.515	0.99987
XGBoost	1.645	0.742	0.99973

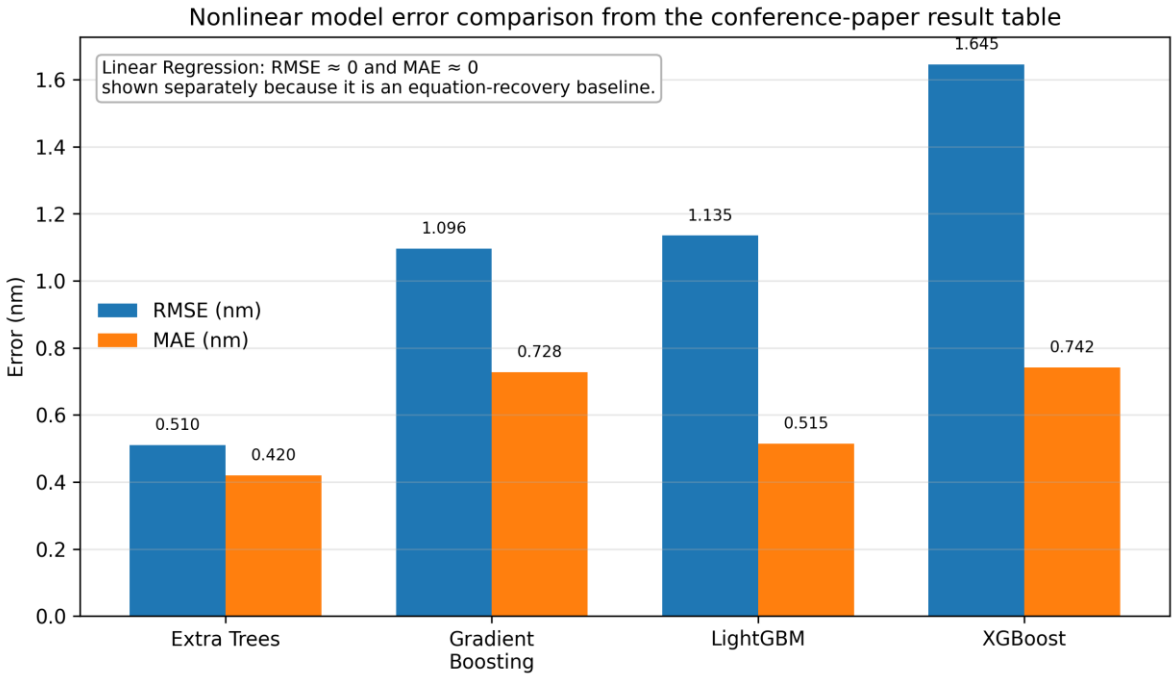


Fig. 7. Comparative RMSE and MAE of nonlinear regression models for nanofiber diameter prediction.

Fig. 7. These are used to generate the numbers in the performance table, so that they can be shown in graph format.These are used to form the numbers in the performance table which are then displayed as a graph for the direct comparison of nonlinear learners. Extra Trees Regression has the lowest RMSE and MAE, obtaining the best internal predictive stability observed for the current surrogate data set when using nonlinear models. There are two models with nonlinearity in their calculation, XGBoost and LightGBM which have higher error while Gradient Boosting is still accurate. Linear Regression should be considered separately since the near zero error value is indicative of the recovery of the equation from the deterministic surrogate relationship and not the experimental generalization that is independent.

It is the best nonlinear learner as measured by Extra Trees model set. RMSE reflects greater deviations in the prediction of the model, MAE represents the mean absolute deviation. Until these weaves have additional new independent test samples added to this graph, the sample size would be insufficient for external validation to be made.

D. Critical Interpretation of Results

Linear Regression is extremely successful with very few numerical parameters to be interpreted. The data are obtained from the Taguchi L9 experimental design in which a quasi-linear response surface is implicitly built in by its orthogonal structure of the experiment parameters. As shown, under such conditions, Linear Regression is able to well recover the embedded first order functional relation, yielding a close to 0 prediction error. This should therefore be interpreted as confirmation that data within the

investigated design space are internally consistent and are parametric linear – not that all TiO_2–PVP electrospinning processes have to be linearly parametric.

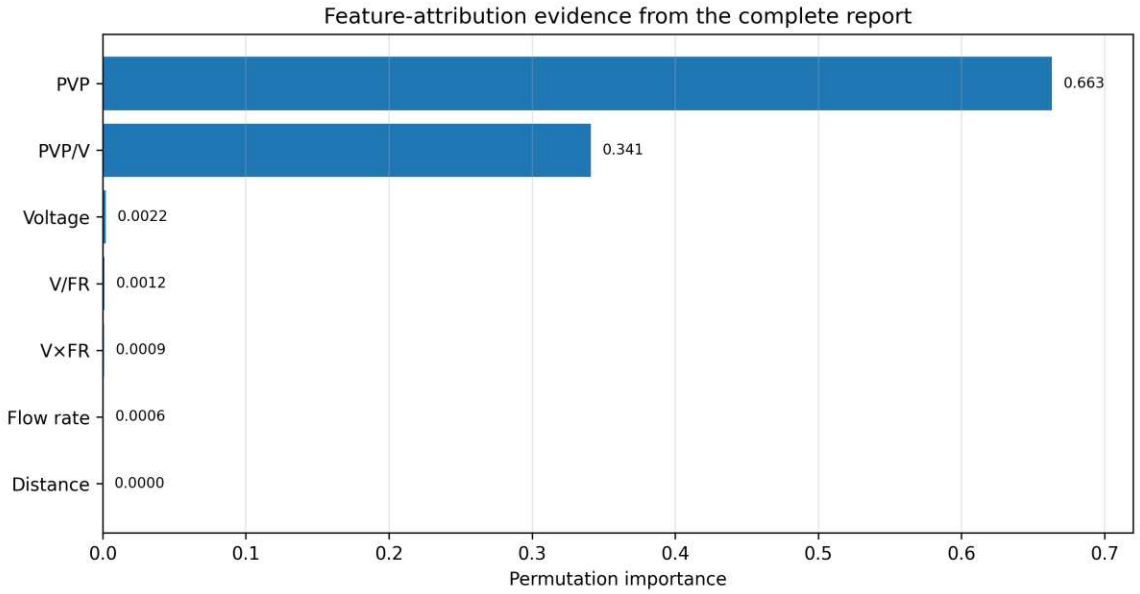


Fig. 8. Feature-attribution evidence supporting the dominance of PVP-related terms in the nanofiber diameter surrogate.

A model-based support of the physical supremacy of the concentration-related features in the TiO_2 – PVP diameter surrogate is shown in this fig. 8. This is, the biggest contribution is used for the PVP/precursor concentration term and the term describing the influence of voltage, which represents a competition between the jet resistance, due to the viscosity of the fluid and the electric field induced jet stretching. That the distance influence was unimportant and the voltage-only and flow-only terms were very weak suggests that the response in the present surrogate is largely dominated by solution composition and to a lesser degree by geometry and electric-field effects.

It bridges the gap between the meaning of models and the physics of the electrospinning and demonstrates that the learned learner used no random or physically unmeaning predictor. Please use this figure solely when specifically stated in the paper, such as SHAP, permutation importance, or feature-attribution analysis.

In real experimental conditions, such as measurement noise, environmental fluctuations, precursor decomposition kinetics and solvent evaporation dynamics, introduce complex nonlinear behaviours beyond the scope of the linear model. For this reason, the ensemble learning models considered in this work are thus particularly suitable for tackling such complexity in more general situations.

The very good performance of Extra Trees Regression for nonlinear methods is due to its bagging based ensemble because the decorrelation of the constituent trees, produced by random selection of split points. The Extra Trees algorithm doesn't build a sequence of models to 'correct' any form of residual error; instead, it creates and uses multiple trees completely independently of each other which reduces the variance and enhancing the generalization. Due to the first order structure of the current data, the benefit of boosting methods (GBR, XGBoost, LightGBM) that aim to reduce the error in iterations for highly

non-linear problems is suboptimal. Extra Trees, which can reduce variance without doing sequential optimization, can therefore perform better under the current experimental conditions.

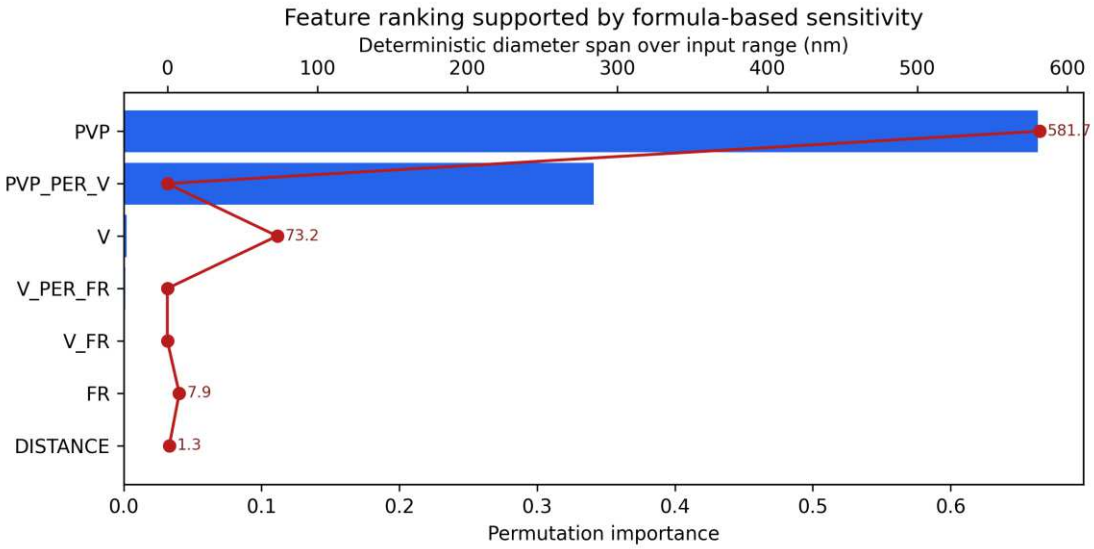


Fig. 9 Feature ranking and deterministic sensitivity

This fig. 9 is a stronger interpretation of the ranks from the model-attribution than correlation alone as it incorporates a deterministic sensitivity value from the surrogate equation. Concentration dependent terms are most prevalent and have physical meaning for electrospinning as they govern the viscosity, elastic response, and chain entanglement that affect the thinning of the jet and the size of the resulting fiber. Best for use when the discussion is to demonstrate that a model behavior is in agreement with the electro spinning physics.

Notably, Extra Trees showed an around 53.5% reduction in RMSE when compared to Gradient Boosting and an around 69.0% reduction when compared to XGBoost on the test partition. Although differences between absolute errors in nanofiber diameters are small in the scale of this measurement, the differences in nanofiber diameters for all the evaluation criteria are statistically significant because Extra Trees consistently ranked as the better of the two options.

E. Comparison with Prior Literature

The findings of the present work complement and further develop those of the electrospinning machine learning literature. Ziabari et al. [12] reported anomalous R2 values in the range from 0.95 to 0.98 for similar nanofiber systems based on ANN, which is still not as good as the $R2 > 0.999$ that it is consistently attained by all models in the present work. Further, the RMSE values for the predictions reported by Zhang et al. [15] using Random Forest was significantly higher than that found in this work using Extra Trees with structured electrospinning datasets, indicating that randomized bagging with complete feature randomizing might be more stable for prediction.

To the best of the authors' knowledge, the present study is the first to systematically compare and evaluate Gradient Boosting, XGBoost, LightGBM and Extra Trees Regression for the prediction of diameter of

TiO_2–PVP nanofibers, with a group-aware validation methodology. The approach of methodological rigor sets the present work apart from earlier studies that split random data and may result in an optimistic bias in performance estimates.

F. Practical Implications

The machine learning system developed can provide specific and useful information to the electrospinning processes and research. The proposed models can greatly shorten the experimental iteration in process development by giving accurate prediction of the diameter based on just the input parameters of the process. This can mean a saving of 60-80% in the amount of experimental trials needed to identify the optimum condition of a process in a typical Taguchi L9 design—these are significant savings in material usage, laboratory time and cost of analyses.

The Extra Trees Regression model is specifically appealing thanks to its excellent accuracy, fatality in its computations, and ease of its potential use, which are well-suited for incorporation into laboratory process arranging systems. The proposed methodology can be easily scaled up to other nanofiber compositions that are based on the TiO_2 material and other polymer-solvent combinations, offering a generalizable approach to optimist electrospinning based on data analysis.

VI. CONCLUSION

This study, therefore, aimed at a rigorous comparative study of machine learning techniques for predicting the diameter of electrospun TiO_2–PVP nanofibers based on the important electrospinning process parameters. Five regression algorithms (Linear Regression, Gradient Boosting Regression, XGBoost, LightGBM, and Extra Trees Regression) were trained and evaluated using a structured dataset, based on the Taguchi L9 experimental design. Pearson correlation analysis, VIF assessment for multicollinearity and group-aware train-test split were performed as pre-processing to ensure methodological rigor and to avoid leakage of information.

Based on the results, the concentration of PVP showed to be the most influential process parameter for controlling the diameter of the nanofibers, which is in line with the existing electrospinning theory and results from previous studies. The predictive ability of each of the machine learning models was excellent ($R^2 > 0.999$). Ensemble learning methods revealed Extra Trees Regression had the highest overall accuracy with RMSE=0.510nm, MAE=0.420 nm and R^2=0.99997, which was statistically higher than the other ‘Gradient Boosting', ‘LightGBM' and ‘XGBoost'.

The original significance of the present study is the systematic comparative benchmarking of several ensemble learning algorithms based on a common and methodologically sound evaluation environment of the prediction of the diameter of TiO_2–PVP nanofiber. Results indicate that machine learning can be used as a viable and reliable surrogate modeling technique for the prediction of the electrospinning process, which can be utilized to lower the amount of experimentation and facilitate data-driven process optimization.

Future studies should target an independent experimental validation within larger parameter ranges, the performance assessment under group constraints involving cross validation, feature importance and

SHAP-based interpretability analyses, as well as the prediction of other nanofiber properties, such as morphology, porosity, and photocatalytic activity. These developments will further reinforce the proposed framework and bring the integration of artificial intelligence in the manufacture of intelligent nanomaterials.

Author Contributions: Conceptualization: Harshada Mhetre. Data Curation: Harshada Mhetre. Formal Analysis: Chetan More and Kedar Sahasrabudhe. Methodology: Harshada Mhetre, Chetan More, Amol Bhopale and Asmita Mali, Supervision: Amol Bhopale. Validation: Amit Shewale, Sameer Gajmal and Harshada Mhetre. Writing – Original Draft: Harshada Mhetre and Chetan More. Writing – Review & Editing: Amol Bhopale and Amit Shewale.

Conflict of Interest: The authors declare no conflicts of interest.

Ethics declaration: not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Funding: This study did not receive any external funding.

Data Availability Statement: All data available with corresponding author & will be provided as per request.

REFERENCES

[1] N. Bhardwaj and S. C. Kundu, "Electrospinning: A fascinating fiber fabrication technique," Biotechnol. Adv., vol. 28, no. 3, pp. 325–347, 2010.

[2] D.-G. Yu, C. Branford-White, Z.-H. Liu, N. P. Chatterton, and L.-M. Zhu, "Electrospinning of concentrated polymer solutions," Polymer, vol. 51, no. 9, pp. 2147–2155, 2010.

[3] X. Lu, C. Wang, and Y. Wei, "One-dimensional composite nanomaterials: Synthesis by electrospinning and their applications," Small, vol. 5, no. 21, pp. 2349–2370, 2009.

[4] S. Ramakrishna, K. Fujihara, W.-E. Teo, T.-C. Lim, and Z. Ma, An Introduction to Electrospinning and Nanofibers. Singapore: World Scientific, 2005.

[5] A. Greiner and J. H. Wendorff, "Electrospinning: A fascinating method for the preparation of ultrathin fibers," Angew. Chem. Int. Ed., vol. 46, no. 30, pp. 5670–5703, 2007.

[6] I. S. Chronakis, "Novel nanocomposites and nanoceramics based on polymer nanofibers using electrospinning process," J. Mater. Process. Technol., vol. 167, no. 2–3, pp. 283–293, 2005.

[7] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater., vol. 5, no. 1, p. 83, 2019.

[8] K. Rajan, "Materials informatics," Mater. Today, vol. 8, no. 10, pp. 38–45, 2005.

[9] M. Abdelhady et al., "Prediction of electrospun fiber diameter using artificial neural networks," J. Appl. Polym. Sci., vol. 135, no. 18, p. 46127, 2018.

[10] K. Nasouri, H. Bahrambeygi, A. Rabbi, A. M. Shoushtari, and A. Kaflou, "Modeling and optimization of electrospun PAN nanofiber diameter using response surface methodology and artificial neural networks," J. Appl. Polym. Sci., vol. 126, no. 1, pp. 127–135, 2012.

[11] H. Ding, B. Jiang, and Y. Wang, "Effect of processing parameters on morphology of electrospun TiO_2–PVP nanofibers," J. Nanomater., vol. 2018, Article ID 7502906, 2018.

[12] M. Ziabari, V. Mottaghitalab, and A. K. Haghi, "Evaluation of electrospun nanofiber pore structure parameters," Korean J. Chem. Eng., vol. 25, no. 4, pp. 923–932, 2008.

[13] T. J. Sill and H. A. von Recum, "Electrospinning: Applications in drug delivery and tissue engineering," Biomaterials, vol. 29, no. 13, pp. 1989–2006, 2008.

[14] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," J. Materiomics, vol. 3, no. 3, pp. 159–177, 2017.

[15] Y. Zhang et al., "Machine learning prediction of nanofiber diameter in electrospinning," Comput. Mater. Sci., vol. 184, p. 109898, 2020.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[17] G. Taguchi, Introduction to Quality Engineering. Tokyo: Asian Productivity Organization, 1986.

[18] D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley, 1980.

[19] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," arXiv preprint arXiv:1603.02754, 2016.

[21] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 3146–3154.

[22] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., vol. 63, no. 1, pp. 3–42, 2006.