Available online at www.sciencedirect.com

Manufacturing Letters

Manufacturing Letters 41 (2024) 364–374

52nd SME North American Manufacturing Research Conference (NAMRC 52, 2024)

Harnessing Interpretable and Ensemble Machine Learning Techniques for Precision Fabrication of Aligned Micro-Fibers

Imtiaz Qavi[a], George Tan[a]*

[a]Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, Lubbock, TX 9409

* Corresponding author. Tel.: +1 806-742-3543; *E-mail address:* george.z.tan@ttu.edu

Abstract

Electrospinning is a robust technique for producing micro/nano-scale fibrous structures, influenced by intricate interplays of fluid dynamics, aerodynamics, and electromagnetic forces. Depending on the desired outcome, these fibers can adopt various morphologies, including solid, tubular, concentric, and gradient. Such morphologies are modulated by parameters such as collector configuration, flow rate, voltage, solution properties, and nozzle dimensions. However, the task of modeling and predicting these multifaceted morphologies remains complex. Aligned microfibers with 3D orientation hold promise in tissue engineering, regenerative medicine, and drug delivery, necessitating meticulous control over the fabrication parameters. In our research, we tapped into machine learning (ML) to address these challenges. Classification ML models were designed to predict fibrous patterns—aligned, random, or jet branching—based on determinants like voltage, flow rate, and collector configurations. Notably, the Random Forest (RF) and Support Vector Machine (SVM) models, especially with radial kernel-trick, displayed outstanding predictive capabilities on the test data. Furthermore, regression-based ML was harnessed to discern fiber alignment coherency and inter-fiber distances. Models such as Lasso and Ridge regression elucidated predictive coefficients for these characteristics, while ensemble models, like gradient-boosting (GB) decision trees (DT), showcased prowess in regression scenarios. Key findings spotlighted the significance of parameters like plate gap for alignment coherency and needle-to-collector distance for inter-fiber spacing. As we strive to gain granular control over micro/nano feature morphology in electrospinning, understanding predictor-response dynamics is imperative. Our investigation underscores the essential role of ML in enhancing both qualitative and quantitative precision in fabricating advanced fibrous structures. Moreover, fusing ML with real-time process monitoring offers groundbreaking potential, particularly in Bio-Fabrication, regenerative medicine, and tissue engineering, where high-precision manufacturing remains a top priority.

Keywords: *Electrospinning; Machine Learning; Support Vector Machines; Random Forest; Gradient Boosting.*

1. Introduction

Electrospinning is a rapid fabrication technique for polymer-based fibrous structures at the micro and nanoscale using dielectric polymer solutions. With applications spanning across diverse domains like healthcare, soft materials, flexible electronics, energy harvesting, chemical storage, and catalysis, its relevance cannot be overstated [1, 2]. The functional utility of these fibrous structures in specific application areas hinges predominantly on their micro and macro attributes. At the microscopic level, properties such as diameter of the fibrous structures, surface porosity, material composition, and hollow cross-sections are of paramount importance [3]. On the

macroscopic scale, attributes like fiber orientation (be it aligned or random), orientation degree, dominant directional flow, inter-fiber spacing, and fiber density play pivotal roles [4, 5]. From a manufacturing standpoint, the meticulous manipulation of process parameters facilitates the creation of fibrous constructs that exhibit the desired micro and macro characteristics. Hollow tubular scaffold structures are increasingly sought after in microfluidics, finding applications in areas such as healthcare, drug delivery, and tissue engineering [6, 7]. When it comes to the fabrication of these hollow, micro-scale fibrous structures (referred to as microtubes), process parameters like flow rate, the distance between needle and collector, and plate-gap critically influence the final product's orientation and fiber density. Electrospun fibrous assemblies with a high degree of alignment have found substantial use in tissue engineering, disease modelling, and studies of cell migration, to name a few [8, 9].

One of the major challenges of fabricating highly aligned micro/nano fibers via electrospinning is its high sensitivity to variations in process parameters. To fully harness control over this fabrication technique, it is crucial to understand how these process parameters (predictors) influence both the micro and macro-outcomes (responses) of our process. Various experimental modeling techniques, such as design of experiments, correlation analysis, principal component analysis, regression, and other statistical and machine learning models, are currently employed to develop predictor-response models for the properties of Electrospun fibrous structures at both the macro and micro scales [10-13].

Supervised machine learning (ML) techniques, such as linear regression, decision trees, support vector machines, and neural networks, have been successfully integrated into the electrospinning fabrication process [13-16]. Pervez *et al.* used partial least squares (PLS) and support vector regression-based ML techniques to predict diameters of chitosan-based nanofibers with a high prediction accuracy [14]. Sarma *et al.* used an ensemble of ML techniques such as gradient boosting (GB), XGboost (XGB), linear regression, and random forest (RF) to determine crucial predictors such as feed, polymer concentration, and Flory-Huggins Chi parameter, and how they affect the polyvinylidene fluoride (PVDF) polymer fiber diameter [15]. Morphological parameters such as diameter and inter fiber separation as well as post-fabrication mechanical properties such as strain at break and ultimate tensile strength for polyvinyl alcohol (PVA) for seven different process parameters were modelled using artificial neural network (ANN) by *Roldán et al.* [17]. Their ML-based modelling approach showed high prediction accuracy and allowed the biomimetic vascular implant fabrication using PVA for their optimal manufacturing setup mined through their ML models [17]. Furthermore, ML techniques are also used to model post-fabrication application of nanofibers in tissue engineering applications. For instance, Sujeeun *et al.* investigated the crucial in-vitro and physio-chemical data of Electrospun polymeric scaffolds applied to skin tissue engineering applications [18]. Their experiment showed that fiber and pore diameter were the most crucial physio-chemical predictors of performance for the synthetic tissues [18]. Hence, ML has shown that it is possible to predict and interpret process parameters that affect the overall electrospinning outcomes. However, these techniques frequently encounter the bias-variance tradeoff, especially in complex model forms [19]. Therefore, selecting an appropriate model that can accurately fit training data with low bias and predict unknown data with low variance is of paramount importance for process modeling.

Another crucial challenge arises when complex machine learning models act as "Black Box" models, where there is no information available about cause-effect relationships based on physical mechanisms despite their high prediction accuracies [20]. This phenomenon is common in complex models such as neural networks, random forests, and various boosting techniques [21]. While simpler models such as regression and decision trees might not always match the prediction accuracy of more complex models, they offer a clear advantage in interpretability [22]. In manufacturing process improvement, it's not just about achieving high predictive accuracy; understanding the underlying relationships between predictor and response variables — essentially how the model derives its predictions — is equally crucial.

In this experiment, our goal was to utilize minimal data from the post-fabrication process to assess qualitative process variations that result in classifiable outcomes such as random/align fibers (from single jets) and multi-jet electrospinning situations. Furthermore, in our effort to fabricate highly aligned microfibers, we employed multiple classification and regression-based machine learning techniques to map process parameters to outcomes with high alignment coherency and density. Classification-based ML enabled us to identify sets of process parameters conducive to forming either aligned or random fibers from single jets and discerning scenarios of jet branching from singular electrospinning jets. For the classification problem, we employed Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forest (RF). For regression predictions on fibrous structure alignment coherency and inter-fiber distance, we used penalized linear regression models like Lasso and Ridge, Decision Trees, Random Forest Regressor, and Gradient Boosting (GB) ensemble techniques. Prior literature indicates that these techniques perform well with small datasets (< 100 data points) [15, 23, 24], unlike deep learning techniques, which achieve accurate predictions but require larger training datasets [25]. Floares *et al.* demonstrated the impact of sample size on the performance of the random forest (RF) technique. Their results showed that prediction accuracy reached a threshold with an area under the curve (AUC) > 95% with a sample size < 100[26]. While these techniques have proven effective under limited data conditions, model training through augmented or synthetic data, using design of experiments (DOE)-based techniques and random noise addition, has been reported to mitigate the limitations associated with small data samples from experimental research [27, 28].

Our findings underscored a trade-off: while simpler models elucidated the relationship between predictors and responses, they sometimes compromised on prediction accuracy. Conversely, sophisticated ensemble models like Random Forest and Gradient Boosting excelled in accuracy. In this research, we melded the strengths of both easily interpretable models, shedding light on influential process parameters, and more intricate techniques known for their predictive prowess. The data-driven conclusions affirm that machine learning is adept at process classification and adeptly predicting unfamiliar data, especially when traditional linear models fall short. This suggests the potential of machine learning as an invaluable tool for in-process diagnostics, enabling real-time parameter adjustments grounded in live learning from accumulated data.

2. Materials and Methods

2.1. Electrospinning System

The electrospinning process was carried out in the core-sheath co-axial mode with the objective of producing tube-like fibrous structures. Poly-Caprolactone (PCL) with a molecular weight of 80,000, acquired from Sigma-Aldrich (St. Louis, MO), was prepared in a 12% (w/v) concentration by mixing it with Dichloromethane (DCM) obtained from VWR (Radnor, PA). Poly-ethylene-oxide (PEO) with a molecular weight of 300,000, also purchased from Sigma-Aldrich, was prepared in a 6% (w/v) solution by mixing it with DCM. The PCL-DCM solution served as the sheath solution, while the PEO-DCM solution was used as the core solution in the co-axial electrospinning system. For our experiment, a robotic electrospinning system (Tongli-Tech, China) was utilized, as depicted in **Figure** 1. An in-house parallel plate collector with an adjustable plate gap was employed for collecting the fibrous structures. The co-axial setup utilized 26 core needles and 21 sheath needles. In this experiment, the core-to-sheath flow rate was maintained at a 1:2 ratio, with the combined volumetric flow rate indicating their bulk flow rate.

2.2. Design of Experiments and DOE-informed data augmentation

In this experiment, we employed four predictor variables: Voltage, Bulk Flow Rate, Needle to Collector Distance, and Plate Gap. A rule of thumb is to use a training dataset that is 10 times larger than the total number of features for training [29-31]. These variables were utilized to predict fiber orientation, alignment coherency, and inter-fiber distance. The design of experiments (DOE) followed a full factorial design [32] with four factors, each having three levels. This setup resulted in a total of 81 data points for analysis. The ranges of the operating parameters for our design space are shown in **Table 1**:

Table 1. Predictor variable bounds for the design space

Parameter	Lower Bound	Upper Bound
Voltage	10 KV	20 KV
Bulk Flowrate	3.5 ml/h	8 ml/h
Plate Gap	2 cm	6 cm
NC distance	10 cm	25 cm

Additionally, 30 additional design points were generated using the Box-Behnken design [33]. These points included six replicates of the center point and were designated for use as our test set data. Due to the imbalanced response from the initial DOE dataset, consisting of 81 data points, we opted for design augmentation to address this issue by manually introducing additional parameter combinations within the existing design space. To rectify the imbalance, we gathered supplementary parameter configurations corresponding to Aligned, -Jet branch, or Random fibers. The original full factorial design table underwent augmentation to incorporate the necessary conditions for achieving a balanced response dataset for the classifiable jet instances. Therefore, a balanced dataset, including equal instances of aligned, random, and branching jet conditions (1:1:1 ratio), was established by augmenting the design table. This involved adding 42 new operating combinations to the existing 81 data points, resulting in a total of 123 row data. The selection of these new operating conditions was done manually to ensure the occurrence of 41 instances each for aligned, random, or jet-branching fibers. Subsequently, the data points obtained from the DOE approach were fitted to a polynomial response-surface model (RSM). The resulting models demonstrated a high multiple-adjusted R-squared value ($R^2 > 0.95$) and an insignificant lack of fit ($P > 0.10$), indicating a robust fit to the augmented design table [34]. These models were then utilized to generate synthetic data using a DOE-informed data augmentation approach [35]. For the classification-type problems, synthetic data were generated based on nominal logistic regression techniques based on predictor-response data obtained using DOE [36-38]. This served as training data for the proposed machine learning models. To expand the dataset, the original 123 rows were sampled based on the DOE-informed RSM model using a Monte Carlo deterministic sampling method [39]. This process yielded an approximate total of 1002 balanced synthetic data points with 334 cases for each class, which were employed for training machine learning techniques.

2.3. Imaging Systems

The images were captured using an Olympus DSX series digital microscope and subsequently processed using ImageJ [40] for getting numeric data about alignment coherency (using AlignmentJ plugin) and inter fiber distance. The aligned and random fibers were determined based on their orientation on the collector (**Figure 2e**) and the Jet branching was determined based on the near-field imaging of the Taylor cone. It is to be mentioned that jet branching yielded fully chaotic instances of microfibers emanating from multiple jets that cannot be classified under random or aligned orientations, hence a separate classification.

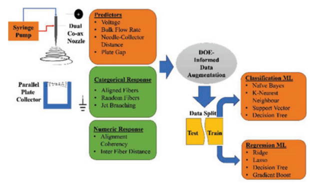

Fig. 1. Schematic of the Electrospinning System Setup and the predictor/ response data flow for machine learning techniques.

The near-field Taylor cone images were captured using a Sigma 300 mm macro lens with a Nikon TC-201 2X teleconverter on a Nikon D3200 DSLR camera. Together with the teleconverter, macro lens, and the DSLR's 1.5x crop factor, a 900 mm equivalent 24.2 megapixels high resolution images

were obtained to distinguish between single, multiple, or no jet formation on the Taylor Cone. The images were used to label the original data set for the classification-type machine learning approach.

2.4. Machine Learning techniques

The machine learning problems of the classification type involved identifying process parameters that led to aligned fibers, random fiber formation, or jet branching from the Taylor Cone [41]. These phenomena were classified into three distinct outcomes for our analysis. To predict these qualitative outcomes, we employed classification models, namely Naïve Bayes, K-Nearest Neighbor, Support Vector Classifier, and Decision Tree. The accuracy of the classification models was calculated based on the total number of true positives and negatives out of the overall observations in the predicted and test sets which forms the confusion matrix [42], as follows:

$$Accuracy = \frac{True\ Positive + Tru\ \ Negative}{Total\ Observations} \tag{1}$$

The naïve bayes classification method is based on the Bayes theorem conditional probability that determines the conditional outcome of event A given event B and is expressed by the following form [43]:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)*P(A)}}{\mathrm{P(B)}} \tag{2}$$

Local learners like the K-Nearest Neighbor classification technique does not construct a specific model; instead, it employs the training data to predict the classification of unknown test data [44]. In this approach, the training set essentially serves as the prediction model [44]. The user specifies the number of neighbors (K) to be considered for classifying the unknown data. The algorithm then computes the Euclidean distance between the test point and neighboring training data, as described by the formula [45]-

$$d = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{3}$$

Here, $(x_i - y_i)$ is the distance between the unknown point in the test dataset and the known point in the training dataset. For our KNN classification model, we used a wrapper function to tune the hyperparameters of the model based on a 10-fold cross validation approach by varying K from 1 to 1000 on the training data set [46].

The support vector machine (SVM) classifies responses by fitting an optimal hyperplane that best separates distinct responses [47]. The data points in the training set closest to this hyperplane, known as marginal points, are termed support vectors [48]. In SVM machine learning, the data are sometimes not linearly separable. In such cases, the Kernel Trick [49] is applied, transforming the data into a higher-dimensional feature space where it can be linearly separable using an n-dimensional hyperplane [50]. The SVM commonly employs linear, polynomial, or radial kernel trick methods to aid classification [51]. The hyperparameters of linear SVM (cost function), polynomial SVM (cost function and degrees), and radial SVM (cost function and gamma) were tuned using a wrapper function based on 10-fold cross-validation. The values varied within the range $22 \leq$ Cost Function ≤ 28, $2 \leq$ Degrees ≤ 8, and $0.001 \leq$ Gamma ≤ 10.

Decision trees split data points continuously based on similarities and differences in features (predictors) [52]. These trees are interpretable through nodes, sub-nodes, and leaves (end points) where data are classified based on separable features [53]. Random Forests and bagging are ensemble learning techniques [54] that enhance single decision tree outputs by averaging decisions from numerous trees, reducing overfitting in the training set. The decision tree was optimized using pruning [55], identifying the optimal complexity parameter (CP) [56] based on minimum error post-fitting, and using this optimal hyperparameter for refitting a new tree.

In regression problems, we began with simple multiple regression techniques, such as Lasso (least absolute shrinkage and selection operator) and Ridge regressions. Ridge regression, or L2 regularization, shrinks model coefficients close to zero based on their influence [57]. Lasso regression (L1 regularization) can set model coefficients to zero depending on their influence [58]. Both Ridge and Lasso regressions use a penalty function or shrinkage parameter (lambda) to shrink the model coefficients [59]. The model forms for Ridge and Lasso regressions are [60]:

$$\text{Ridge Regression Model: } RSS + \alpha \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

$$\text{Lasso Regression Model: } RSS + \alpha \sum_{j=1}^{p} |\beta_j| \tag{5}$$

Here, RSS is the residual sum of squares from an ordinary least square (OLS) fit, p= number of predictors, β = shrinkage or the penalty parameter. The optimal lambda was tuned using 10-fold cross validation for glmnet (version 4.1-8).

For the regression problem, the Decision Tree regressor was pruned using the optimal complexity parameter as in the classification type problem. The Gradient Boosting (GB) regressor uses error information from the prior trees to improve the model for n number of tree constructs [61]. The learning rate is a crucial parameter in this regard, and we tuned it with a range of learning rates from 10^{-10} to 10^{1} and extracted the best r-squared value in the test set using a loop function.

The r-square values for the regression models were calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

$$Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{7}$$

$$R^2 = 1 - \frac{MSE}{Var(y)} \tag{8}$$

Where, MSE = mean squared error, Var(y) = variance, y_i = actual response, $\hat{y}_i$ = predicted response, $\bar{y}$ = sample mean.

3. Results and Discussions

3.1. Microscopic and Near-Field Images

The data concerning jet branching and single-jet electrospinning was documented during the fabrication process. The single or branched jetting methods produced either aligned fibers collected between two parallel plates or random fibers gathered in the space between these plates. **Figure 2** displays dark-field images captured with an Olympus DSX microscope from single or multiple jet systems, showcasing aligned and random. No fiber conditions (**Figure 2e**) occurred outside our design space depicted in **Table 1**. The inset in **Figure 2** shows formation of tubular micro fibrous structure from the selected core: sheath configuration after washing away the PEO core in water. It is to be mentioned that jet branching yielded a full-chaotic scenario where end fiber could not be classified into either random or aligned, hence a fully different class was allotted to it. The Taylor cone near-field images are presented in **Figure 3**. The single-jet condition as depicted in **Figure 3a** and **3b** both the single and branched-jet conditions resulted in aligned and random fiber orientations. It is to be mentioned that the no-fiber condition (**Figure 2d**) resulted from the no jet Taylor cone instance (**Figure 3c**) and did not appear in any experimental run within our design space. Hence, the classification problems were mainly based on aligned, random, and jet-branching; no-fiber/ no-jet conditions were not present in our dataset.

3.2. Naïve Bayes (NB) Classifier

The Gaussian Naïve Bayes classifiers compute the a-posterior probabilities using Bayes' rule [62]. For each class, a probability distribution plot (**Figure 4**) is generated based on the predictors. Our model output plot shows that jet branching is more prominent (higher probability) for large voltages (>15 kV) and shorter distances (<18 cm) with a moderate plate gap (4 cm) and combined flow rate (6 mL/h). Aligned fibers exhibit higher production probabilities for moderate voltage (~15 kV), moderate needle-collector distance (18 cm), and a moderate plate gap and flow rate, as depicted in **Figure 4**. Randomly formed fibers have a higher probability at lower voltages (<14 kV), shorter needle-collector distances (<16 cm), larger plate gaps (>5 cm), and lower flow rates (<6 mL/h). The Naïve Bayes classification achieved a high accuracy of 92% on our test set. Although the Naïve Bayes classification algorithm requires feature independence [63], it has widely been used in many experimental modeling applications with high prediction accuracies [64-67]. Also, in our application, the descriptive features, namely voltage, flow rate, plate gap, and needle-collector distance, can be considered independent by the theory of probability [68] if:

$$\mathrm{P}(X_1 | X_2) = \mathrm{P}(X_1) \tag{9}$$
$$\mathrm{P}(X_1, X_2) = \mathrm{P}(X_1) \times \mathrm{P}(X_2) \tag{10}$$

where, X_1 represents the probable value of the feature variables (such as voltage, distance, plate-gap, and needle-collector-distance) given the value of another feature variable (X_2) apart from itself, it is notable that in our application, the value of one feature variable does not depend on the value set for another feature variable.

3.3. K-Nearest Neighbor (KNN) Classifier

In contrast, the KNN technique lacks interpretability like the Naïve Bayes approach. When applied to the test data, the KNN approach exhibited a lower classification accuracy (53%). Therefore, directly predicting the test set using the KNN method, based on the training set data, did not prove to be useful for our specific multiclass dataset. The classic KNN models suffer from such skewed training datasets and the classifications is heavily dominated by the majority class during the test-set predictions [69]. This was avoided in our case by balancing both the original data set and the augmented data set with an equal number of cases for the training data. Misclassification and sensitivity to outliers are major issues for non-optimal neighbor size selection in the KNN algorithm because it employs a distance-based prediction method [70-73]. Classification on the test set ignores the closeness between the data points and makes predictions based on the majority vote in the conventional KNN algorithm [73]. Furthermore, the curse of dimensionality is another common issue leading to low accuracy in the KNN algorithm, as it results in inefficient response predictions when the number of feature variables increases [74].

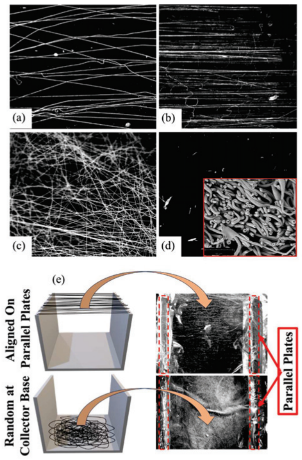

Fig. 2. Qualitative dark-field microscopic imaging of different fiber orientations obtained from the experiment (a) aligned fibers produced without jet branching (b) aligned fibers produced from jet branching (c) randomly oriented fibers (d) no fibers formed or collected between the parallel plates. (e) Formation of aligned fibers between the parallel plates and random fiber at the bottom of the collector. Inset on Figure(d) shows the SEM image of tubular micro fibrous structure formed.

3.4. Support Vector Machine (SVM) Classifier

The SVM approach with a radial kernel (SVM-R) demonstrated the highest accuracy on the test set compared to the linear (SVM-L) or polynomial kernels (SVM-P), with an optimal cost function (= 4) and gamma (= 0.189). Therefore, for classifying multiclass objects such as in our case

(branching, aligned, and random), the radial kernel SVM is the preferable choice for building accurate models. Although

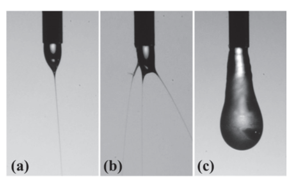

Figure x: Near-field Taylor-cone images for (a) Single Jet (Aligned Fibers) (b) Branched Jet (Aligned and Random Fibers) (c) No Jet (No Fiber).

obtaining interpretable models. Although obtaining interpretable model coefficients directly from an SVM model is not possible, we created visual phase plots for the aligned, random, and branching jet cases by plotting the predictors NC-distance, plate gap, and flowrate against the predictor voltage. In each of the two parameter selections, the other two were set at their midpoint values, i.e., NC-distance = 17.5, Flowrate = 6 mL/h, and Plate Gap = 4.5 cm. These plots were constructed using the radial kernel with the optimal hyperparameter values

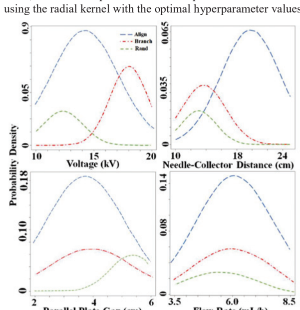

Fig. 4. The probability distribution of the predictors using the Gaussian Naïve-Bayes classifier.

that yielded the best results in our study. In **Figure 5**, the flowrate vs voltage plot shows that for 16 < voltage < 20 kV for flowrates between 4 to 8 ml/h draws the boundary between the formation of Jet branching and Single jets (yielding aligned or random fibers). A narrow region where random fibers were generated was observed between the jet branching and aligned fiber formation zones. The NC-distance vs voltage plot indicated that when the plate gap was 4 cm and flowrate was 6 mL/h, NC-distance < 12 resulted in random fibers for voltages < 14 kV. Additionally, random fibers were observed for plate gaps larger than 5 cm and voltages < 15 kV, as shown in the plate gap vs voltage plot in **Figure 5**.

3.5. Decision-Tree (DT) and Random Forest (RF) Classifiers

The optimal decision tree, with the appropriate hyperparameters, is depicted in **Figure 6**. The first split is based on whether the nc-distance is smaller or larger than 18 cm. It is observed that jet branching primarily occurs when the voltage is > 17 kV, especially for needle-collector distances < 18 cm, and when the voltage surpasses 19 kV for needle-collector distances less than 21 cm. Therefore, both voltage and needle-collector distance emerged as crucial indicators of jet branching phenomena, as revealed by the decision tree analysis. Moreover, the distinction between random and aligned fiber formation seems to be influenced by the needle-collector distance and plate gap when the voltage is < 16 kV. The tree classifies that for a large plate gap (> 5 cm) and a nc-distance < 18 cm, random fibers are formed. The optimal decision tree exhibited an 89% prediction accuracy on the test data. This accuracy was further improved through the application of Random Forest, an ensemble technique that utilizes bootstrap aggregation [75] to combine outcomes from multiple trees based on majority voting or averaging in classification problems [76].

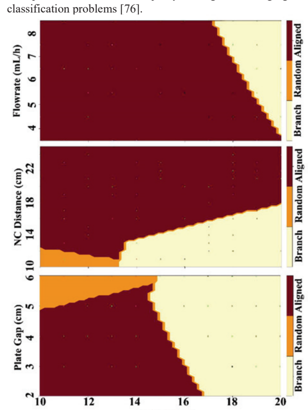

Fig. 5. Phase-Plot of the classification problem using a radial kernel support vector machine (SVM).

Despite the high prediction accuracy offered by the Random Forest model, it lacks interpretability and operates as a "Black

Box" model, providing limited insight into the underlying factors influencing the outcomes.

The formation of random, aligned, or jet branching phenomena using various classification-type machine learning model for our microfiber dataset has been tabulated in **Table 2**. It is observed that Naïve Bayes, Support Vector machines with Linear, and radial kernels showed prediction accuracies > 90% for our interpretable models. Ensemble techniques like RF also showed high prediction accuracies, although no efforts were made in the current experiment to decipher the black-box ensemble technique.

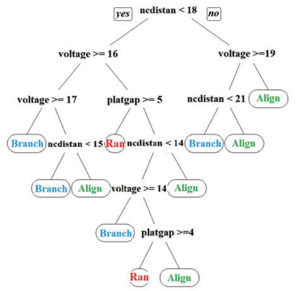

Fig. 6. Decision Tree Classification Aligned, Branch, and Ran (Random) Fiber based on different process parameters. Ran = Random Fiber

Table 2. Classification Machine Learning Summary

Technique	Additional Parameter information	Prediction Accuracy
NB	Null*	94%
KNN	Tuned Neighbour Size (k =12)	53%
SVM-L	Linear Kernal	94%
SVM-P	Polynomial Kernal	74%
SVM-R	Radial Kernal	97%
DT	Pruned (CP = 0.01)	93%
RF	N = 500 Decision Trees	98%

****Null**: No additional parameters*

3.6. Pearson Correlation for collinearity determination

The Pearson Correlation Coefficient (r) indicates the linear correlation between two variables. In regression-based machine learning techniques, it provides insights into choosing suitable models (either linear or non-linear fitting) for a given dataset. **Figure 7** displays the Pearson Correlation matrix. It is noticeable that Alignment Coherency (Align) exhibits a significant positive correlation with nc-distance (named 'Dist' in **Figure 7**) and voltage. It also shows a significant negative correlation with the plate gap (P.Gap). Additionally, the inter-fiber distance (Fib.D) displays a notable negative correlation with bulk flowrate (Flow). Consequently, we employed both linear machine learning models such as Lasso and Ridge regression, as well as non-linear model fits like Random Forest (RF) and Gradient Boosting. The proper choice of linear versus non-linear regressor machine learning models are crucial for numerical data as it may lead to higher or lower prediction accuracies of numerical data based on the chosen model type.

3.7. Ridge & LASSO regressor

Table 3 provides a summary of the intercept and predictor coefficients for the linear machine learning techniques (Ridge and Lasso) applied to our alignment coherency and inter-fiber distance data. In the alignment coherency data, the ridge regression model coefficients exhibit smaller values compared to those of the inter-fiber distance. The plate gap exerts the most significant influence on alignment coherency, displaying an inverse relationship. Hence, within the tested plate gap range (2 - 6 cm), smaller plate gaps lead to better alignment coherency. Furthermore, bulk flowrate has the least impact on the alignment model, as indicated by the ridge regression coefficient. The optimal penalty lambda ('Optm Lambda' in **Table 3**) sets the bulk flowrate coefficient to zero (indicated by x on **Table 3**) due to its negligible effect on the alignment coherency lasso model. In contrast, for the inter-fiber distance model, the bulk flowrate has the most significant effect. With an increase in flowrate, the bulk volume of formed fibers rises, subsequently reducing the inter-fiber distance due to higher fiber density. Both the Ridge and Lasso regression models for the alignment coherency data exhibit high R^2 values (~81%), whereas those for the inter-fiber distance show only 55%. This discrepancy can be attributed to the non-linear relationship between the predictors and the inter-fiber distance, as demonstrated in the Pearson correlation matrix in **Figure 7**.

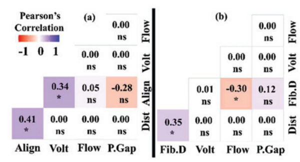

Fig. 7. Pearson's Correlation Table for (a) alignment coherency (%) (Align) and (b) Inter-Fiber Distance (Fib.D) in relation to the different predictors.

3.8. Decision Tree and Random Forest (RF) regressor

Given the non-linear relationship between the predictors and the response variables, we turned to employing non-linear machine learning models, such as decision trees. The decision tree, optimized with the appropriate CP, achieved R^2 values of 71% and 64% on the training set data. The decision tree regressor resulted in intricate tree structures, as illustrated in **Figure 8**.

In our regression-type problems, single decision trees didn't perform well on our test data, despite fitting the training data very effectively. Decision trees exhibit this bias-variance trade-off issue in regression problems, displaying low bias in the training set but resulting in higher variance in test sets. To enhance the efficiency of our regression model, we opted for Random Forest (RF) with n = 1000. Subsequently, it

demonstrated slightly improved R^2 values for both the alignment coherency and inter-fiber distance data, as indicated in **Table 4**.

Table 3. Ridge and Lasso Regression model co-efficient.

Predictor	Ridge (alignment coherency)	Lasso (alignment coherency)	Ridge (inter fiber)	Lasso (inter fiber)
Optm Lambda	0.0041	0.0003	1.84	0.081
(Intercept)	0.3865	0.4191	152.285	126.552
Bulk Flowrate	0.0024	x	-10.4783	-32.4854
Plate Gap	-0.0332	-0.0312	5.6418	12.254
NC Distance	0.01114	0.0107	1.3911	3.9987
Voltage	0.01644	0.0153	6.2760	16.2154

3.9. Gradient Boost (GB) regressor

To enhance our prediction accuracy using decision tree variants, we turned to an ensemble method, Gradient Boosting. This approach constructs the next best model by combining previous models, aiming to minimize prediction errors. Boosting not only decreases model bias but also enhances generalization for predicting unknown datasets, exhibiting lower variance when compared to methods like RF and other ensemble techniques. The key lies in determining an optimal learning rate that prevents the model from overfitting the training data. In our model, we optimized the learning rate using a looped function, varying the rate from 10^{-10} to 1, and minimizing the R^2 on the test set for each iteration. The optimized Gradient Boosting models achieved high R^2 values (~80%) for both the alignment and inter-fiber distance test datasets.

The figure displaying the relative influence of predictors in reducing the mean square error (MSE) in the training dataset for the Gradient Boosting (GB) model is presented in **Figure 9**. The relative influence assigns standardized numeric values on a scale of 1-100 to each of the predictors such that their cumulative equals to 100 [77]. Predictors that are more crucial in reducing the MSE in the training models are assigned a higher relative influence over others [77]. Plate Gap (P.Gap) and needle-to-collector distance (Dist) appeared to exert the most significant influence on the alignment coherency and inter-fiber distance models, as depicted in **Figure 9a** and **9b**, respectively. Therefore, for our regression analysis, the ensemble method using Gradient Boosting demonstrated the most accurate predictions on the test set. **Figures 9c** and **9d** illustrate the comparison between actual and predicted data for the GB model in alignment coherency and inter-fiber distance, respectively.

Table 3 includes a summary of the machine learning models utilized, along with their corresponding R^2 values and parameter optimization details. While linear regression models offered interpretability, their predictive accuracy was limited due to the nonlinear correlation between predictors and responses. The Decision Tree technique exhibited low bias, resulting in a good training set R^2, but suffered from high variance, leading to a poor test set R^2. To enhance the bias-variance trade-off [78], the incorporation of ensemble techniques, such as gradient boosting, improved test set accuracy, albeit at the expense of reduced model interpretability.

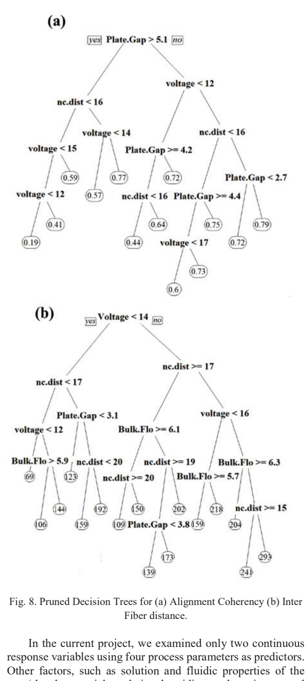

Fig. 8. Pruned Decision Trees for (a) Alignment Coherency (b) Inter Fiber distance.

In the current project, we examined only two continuous response variables using four process parameters as predictors. Other factors, such as solution and fluidic properties of the core/sheath materials, relative humidity, and environmental temperature, were not included in the modeling process. Additionally, many other essential macro and micro parameters, including fiber diameter, cross-sectional hollow portion for microtubular structures, and fiber surface pore composition, have not been modeled using machine learning approaches.

3.10. Discussion and Future Scope

The present paper delves into the applications of interpretable machine learning techniques for the classification and regression modeling of the electrospinning technique. While black-box models such as Neural Networks, Long-

Short-Term Memory (LSTM), and recurrent neural networks (RNN), etc. may offer more accurate predictions, relying solely on these models does not inherently provide an explanation of how the predictions are made. Nowadays, various model-agnostic interpretability techniques are employed to unveil the workings of these black-box techniques, aiming to combine high accuracy with model interpretability. Partial dependency (PD) plots serve as an example of such techniques, where individual predictors are perturbed to observe their effects on the response variable [79]. However, PD plots assume independence among inputs [80], may generate unrealistic scenarios through predictor perturbation [81], and fail to consider interactions between predictor variables. Similar to PD plots, individual conditional expectations (ICE) plots decode black-box models based on data perturbations [82]. While PD plots aim to average the effect of a feature across the entire dataset, ICE plots analyze the effect of a single feature on one observation [83]. Shapley values represent a refinement over PD plots, incorporating feature interactions and avoiding unrealistic predictor settings to provide a more realistic portrayal of black-box models [84]. Local interpretable model-agnostic explanations (LIME) offer another approach to interpreting black-box models by providing local interpretations using surrogate models such as general linear models (GLM) and LASSO [85]. Hence, the surrogate models elucidate the black-box model within the local region. Tree-based surrogate models, such as decision trees, are also utilized to interpret black-box models by identifying feature importance through node splitting [86]. In the future, our objective is to develop models based on artificial neural networks (ANN) and provide interpretability for these constructed models using model-agnostic interpretation techniques.

Table 4. Regression Machine Learning Summary

Technique	Parameter info	Alignment R^2	Fiber Dist R^2
RIDGE	Tuned Lambda	81%	55%
LASSO	Tuned Lambda	80%	59%
DT	Pruned (CP = 0.001)	71%	64%
RF	nTrees = 1000	76%	74%
GB	(lambda = 0.67)	81%	82%

In future investigations, we intend to employ a more comprehensive set of predictors to forecast the outcomes of the electrospinning process across a wider range of response parameters. The experiments will include more categorical measures based on the jetting and fiber collection methods as well as numeric measures for regression modelling. We also intend to use deep-learning techniques with larger datasets on top of the existing machine learning techniques. The deep learning techniques will be used with the model-agnostic interpretability techniques to combine high prediction accuracies with model interpretability that will ensure proper and educated use of black box machine learning models. Additionally, we plan to utilize an in-situ process monitoring system to gather qualitative process information related to the electrospinning jet, including parameters like whipping length, branching distance, and the number of whips. This data will enable us to create a more accurate model for predicting the final fibrous structure outcome. The goal of using fibrous micro-tubular structure in future experiments will be to aid in nutrient and oxygen transport to and removal of waste and biochemical byproducts from tissue culture in 3D micro-environments in hydrogel scaffolds.

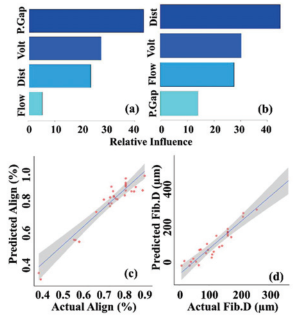

Fig. 9. The relative influence plots obtained from the gradient boosting random forest for (a) alignment coherency (%) and (b) inter fiber distance (µm). Actual vs predicted plots for the (c) alignment coherency (%) and (d) inter fiber distance (µm) from the gradient boosting decision tree regression.

4. Conclusion

In this study, we leveraged ML techniques to model both categorical and numerical responses in electrospinning, based on diverse predictors. The work emphasized the use of traditional machine learning techniques that combine the advantages of good prediction accuracy, interpretability, and training on limited experimental data. Classification algorithms, including Naïve Bayes, Support Vector Classifier, and Random Forest, showcased remarkable accuracy in classifying electrospinning outcomes such as aligned fibers, random fiber formations (both emanating from single jets), and jet branching within the test dataset. Gaussian Naïve Bayes revealed that high voltage and shorter needle-collector distances predominantly instigated jet branching. Concurrently, larger plate gaps were pinpointed by the Naïve Bayes method as key determinants for random fiber formation. The Support Vector Classifier, utilizing an optimally tuned radial kernel, emerged as the most adept model for classification tasks. Its qualitative insights aligned with those from Naïve Bayes, emphasizing larger plate gaps and shorter spinning distances as pivotal for generating random fibrous structures. Decision tree models highlighted voltage values exceeding 17 kV as essential indicators for jet branch formation.

In terms of regression, Ridge and Lasso models indicated a direct relationship between needle-collector distance and voltage with alignment coherency, while an inverse relationship was observed with the plate gap. Conversely, bulk flow rate inversely correlated with inter-fiber distance. Linear models effectively captured alignment coherency data, showcasing strong linear correlations with predictors via the

Pearson Correlation. However, predictions for inter-fiber distance were suboptimal due to inherent non-linearities. Incorporating non-linear regressors, such as the Decision Tree Regressor, led to a modest enhancement in predictive performance. Remarkably, ensemble methods like Random Forest and Gradient Boosting consistently delivered the highest prediction accuracies across both classification and regression tasks. This research accentuates the efficacy of ML in accurately classifying and predicting outcomes in electrospinning micro-fibrous structures. Looking ahead, we aim to integrate ML with in-situ process monitoring, enabling real-time data acquisition for continuous training and optimization of the fabrication process.

Acknowledgement

This work was supported by the National Science Foundation under grant number CMMI-2145108.

References

[1] Li, S., et al., *Hierarchically structured electrospinning nanofibers for catalysis and energy storage.* Composites Communications, 2019. **13**: p. 1-11.
[2] Liu, Z., S. Ramakrishna, and X. Liu, *Electrospinning and emerging healthcare and medicine possibilities.* APL bioengineering, 2020. **4**(3).
[3] Sun, B., et al., *Electrospun anisotropic architectures and porous structures for tissue engineering.* Journal of Materials Chemistry B, 2015. **3**(27): p. 5389-5410.
[4] Pedicini, A. and R.J. Farris, *Mechanical behavior of electrospun polyurethane.* Polymer, 2003. **44**(22): p. 6857-6862.
[5] Meng, X., S.L. Perry, and J.D. Schiffman, *Complex coacervation: Chemically stable fibers electrospun from aqueous polyelectrolyte solutions.* ACS Macro Letters, 2017. **6**(5): p. 505-511.
[6] Wang, G., et al., *Microfluidics-based fabrication of cell-laden hydrogel microfibers for potential applications in tissue engineering.* Molecules, 2019. **24**(8): p. 1633.
[7] Daniele, M.A., et al., *Microfluidic strategies for design and assembly of microfibers and nanofibers with tissue engineering and regenerative medicine applications.* Advanced healthcare materials, 2015. **4**(1): p. 11-28.
[8] Eslamian, M., et al., *Electrospinning of highly aligned fibers for drug delivery applications.* Journal of Materials Chemistry B, 2019. **7**(2): p. 224-232.
[9] Wang, H.B., et al., *Varying the diameter of aligned electrospun fibers alters neurite outgrowth and Schwann cell migration.* Acta biomaterialia, 2010. **6**(8): p. 2970-2978.
[10] Abdelhakim, H.E., et al., *Electrospinning optimization of Eudragit E PO with and without chlorpheniramine maleate using a design of experiment approach.* Molecular Pharmaceutics, 2019. **16**(6): p. 2557-2568.
[11] Khatti, T., H. Naderi-Manesh, and S.M. Kalantar, *Application of ANN and RSM techniques for modeling electrospinning process of polycaprolactone.* Neural Computing and Applications, 2019. **31**: p. 239-248.
[12] Younes, K., et al., *The application of principal component analysis (PCA) for the optimization of the conditions of fabrication of electrospun nanofibrous membrane for desalination and ion removal.* Membranes, 2021. **11**(12): p. 979.
[13] Toscano, J.D., et al. *A machine learning approach to model the electrospinning process of biocompatible materials*. in *International Manufacturing Science and Engineering Conference*. 2020. American Society of Mechanical Engineers.
[14] Pervez, M.N., et al., *Prediction of the diameter of biodegradable electrospun nanofiber membranes: An integrated framework of taguchi design and machine learning.* Journal of Polymers and the Environment, 2023: p. 1-17.
[15] Sarma, S., et al., *Towards an interpretable machine learning model for electrospun polyvinylidene fluoride (PVDF) fiber properties.* Computational Materials Science, 2022. **213**: p. 111661.
[16] Ieracitano, C., et al., *A novel automatic classification system based on hybrid unsupervised and supervised machine learning for electrospun nanofibers*. IEEE/CAA Journal of Automatica Sinica, 2020. **8**(1): p. 64-76.
[17] Roldán, E., et al., *Towards the ideal vascular implant: Use of machine learning and statistical approaches to optimise manufacturing parameters*. Frontiers in Physics, 2023. **11**: p. 1112218.
[18] Sujeeun, L.Y., et al., *Correlating in vitro performance with physico-chemical characteristics of nanofibrous scaffolds for skin tissue engineering using supervised machine learning algorithms*. Royal Society Open Science, 2020. **7**(12): p. 201293.
[19] Singh, S.K., N. Rai, and A. Subramanian, *Machine Learning-Informed Predictive Design and Analysis of Electrohydrodynamic Printing Systems.* Advanced Engineering Materials, 2023. **25**(19): p. 2300740.
[20] Lampathaki, F., et al. *Moving from 'black box'to 'glass box'Artificial Intelligence in Manufacturing with XMANAI*. in *2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. 2021. IEEE.
[21] Guidotti, R., et al., *A survey of methods for explaining black box models.* ACM computing surveys (CSUR), 2018. **51**(5): p. 1-42.
[22] Molnar, C., *Interpretable machine learning*. 2020: Lulu. com.
[23] Nurwaha, D. and W. XJGJTO, *Modeling and Prediction of Electrospun Fiber Morphology using Artificial Intelligence Techniques*. Global Journal of Technology & Optimization, 2019. **10**(1): p. 237-243.
[24] Nurwaha, D. and X. Wang, *The use of adaptive neuro-fuzzy inference systems and support vector machines techniques for evaluation of electrospun nanofiber diameter.* Journal of Computational and Theoretical Nanoscience, 2013. **10**(3): p. 637-642.
[25] Serin, G., et al., *Review of tool condition monitoring in machining and opportunities for deep learning.* The International Journal of Advanced Manufacturing Technology, 2020. **109**: p. 953-974.
[26] Floares, A., et al., *The smallest sample size for the desired diagnosis accuracy.* Int J Oncol Cancer Ther, 2017. **2**: p. 13-19.
[27] Cao, B., et al., *How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics.* ACS nano, 2018. **12**(8): p. 7434-7444.
[28] Fontana, R., et al., *Design of experiments and machine learning with application to industrial experiments*. Statistical Papers, 2023: p. 1-24.
[29] Alwosheel, A., S. van Cranenburgh, and C.G. Chorus, *Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis*. Journal of choice modelling, 2018. **28**: p. 167-182.
[30] Haykin, S., *Neural networks and learning machines, 3/E*. 2009: Pearson Education India.
[31] Baum, E. and D. Haussler, *What size net gives valid generalization?* Advances in neural information processing systems, 1988. **1**.
[32] Anindyajati, A., P. Boughton, and A.J. Ruys, *Study on Processing Parameters of Polycaprolactone Electrospinning for Fibrous Scaffold using Factorial Design.* Regenerative Engineering and Translational Medicine, 2021: p. 1-13.
[33] Konwarh, R., et al., *Diameter-tuning of electrospun cellulose acetate fibers: A Box–Behnken design (BBD) study.* Carbohydrate polymers, 2013. **92**(2): p. 1100-1106.
[34] Goos, P. and B. Jones, *Optimal design of experiments: a case study approach*. 2011: John Wiley & Sons.
[35] Chen, C., et al., *Design of experiments informed deep learning for modeling of directed energy deposition process with a small-size experimental dataset.* Materials & Design, 2022. **222**: p. 111098.
[36] Shah, S., D. Gandhi, and J. Kothari. *Machine learning based Synthetic Data Generation using Iterative Regression Analysis*. in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2020. IEEE.
[37] Rankin, D., et al., *Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing.* JMIR medical informatics, 2020. **8**(7): p. e18910.
[38] Serré, L. and M. Amyot-Bourgeois. *An application of automated machine learning within a data farming process*. in *2022 Winter Simulation Conference (WSC)*. 2022. IEEE.
[39] Fang, S.-E., W.-X. Ren, and R. Perera, *A stochastic model updating method for parameter variability quantification based on response surface models and Monte Carlo simulation.* Mechanical Systems and Signal Processing, 2012. **33**: p. 83-96.
[40] Collins, T.J., *ImageJ for microscopy.* Biotechniques, 2007. **43**(S1): p. S25-S30.
[41] Fernández de La Mora, J., *The fluid dynamics of Taylor cones.* Annu. Rev. Fluid Mech., 2007. **39**: p. 217-243.
[42] Xia, B., et al., *PETs: a stable and accurate predictor of protein-protein interacting sites based on extremely-randomized trees.* IEEE transactions on nanobioscience, 2015. **14**(8): p. 882-893.
[43] Joyce, J., *Bayes' theorem.* 2003.
[44] Garcia, E.K., et al., *Completely lazy learning.* IEEE Transactions on Knowledge and Data Engineering, 2009. **22**(9): p. 1274-1285.

[45]Nayak, S., et al. *Study of distance metrics on k-nearest neighbor algorithm for star categorization*. in *Journal of Physics: Conference Series*. 2022. IOP Publishing.

[46]Sandhu, H., R.N. Kumar, and P. Garg, *Machine learning-based modeling to predict inhibitors of acetylcholinesterase.* Molecular Diversity, 2022. **26**(1): p. 331-340.

[47]Ding, S., X. Hua, and J. Yu, *An overview on nonparallel hyperplane support vector machine algorithms*. Neural computing and applications, 2014. **25**: p. 975-982.

[48]Jakkula, V., *Tutorial on support vector machine (svm).* School of EECS, Washington State University, 2006. **37**(2.5): p. 3.

[49]Murty, M., et al., *Kernel-based SVM.* Support vector machines and perceptrons: learning, optimization, classification, and application to social networks, 2016: p. 57-67.

[50]Tan, Y. and J. Wang, *A support vector machine with a hybrid kernel and minimal Vapnik-Chervonenkis dimension.* IEEE Transactions on knowledge and data engineering, 2004. **16**(4): p. 385-395.

[51]Prajapati, G.L. and A. Patle. *On performing classification using SVM with radial basis and polynomial kernel functions*. in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*. 2010. IEEE.

[52]Gama, J., R. Fernandes, and R. Rocha, *Decision trees for mining data streams.* Intelligent Data Analysis, 2006. **10**(1): p. 23-45.

[53]Wu, C.-C., et al., *Decision tree induction with a constrained number of leaf nodes.* Applied Intelligence, 2016. **45**: p. 673-685.

[54]Dietterich, T.G., *Ensemble learning.* The handbook of brain theory and neural networks, 2002. **2**(1): p. 110-125.

[55]Mingers, J., *An empirical comparison of pruning methods for decision tree induction.* Machine learning, 1989. **4**: p. 227-243.

[56]Ordyniak, S. and S. Szeider. *Parameterized complexity of small decision tree learning*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021.

[57]Wu, D. and J. Xu, *On the Optimal Weighted $\ell_2 $ Regularization in Overparameterized Linear Regression.* Advances in Neural Information Processing Systems, 2020. **33**: p. 10112-10123.

[58]Vidaurre, D., C. Bielza, and P. Larranaga, *A survey of L1 regression.* International Statistical Review, 2013. **81**(3): p. 361-387.

[59]Tibshirani, R., *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996. **58**(1): p. 267-288.

[60]James, G., et al., *An introduction to statistical learning*. Vol. 112. 2013: Springer.

[61]Zhang, Y. and A. Haghani, *A gradient boosting method to improve travel time prediction.* Transportation Research Part C: Emerging Technologies, 2015. **58**: p. 308-324.

[62]Furukawa, D., A. Shimizu, and H. Kobatake, *Automatic liver segmentation method based on maximum a posterior probability estimation and level set method.* 3D Segmentation in The Clinic: A Grand Challenge, 2007. **117**.

[63]Rish, I. *An empirical study of the naive Bayes classifier*. in *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001.

[64]Addin, O., et al., *A Naïve-Bayes classifier for damage detection in engineering materials.* Materials & design, 2007. **28**(8): p. 2379-2386.

[65]Addina, O., S. Sapuan, and M. Othman, *A naïve-bayes classifier and f-folds feature extraction method for materials damage detection.* International Journal of Mechanical and Materials Engineering, 2007. **2**(1): p. 55-62.

[66]Kotzabasaki, M., et al., *Machine learning methods for multi-walled carbon nanotubes (MWCNT) genotoxicity prediction.* Nanoscale Advances, 2021. **3**(11): p. 3167-3176.

[67]Niendorf, K. and B. Raeymaekers, *Using supervised machine learning methods to predict microfiber alignment and electrical conductivity of polymer matrix composite materials fabricated with ultrasound directed self-assembly and stereolithography.* Computational Materials Science, 2022. **206**: p. 111233.

[68]Naik, D.L. and R. Kiran, *Naïve Bayes classifier, multivariate linear regression and experimental testing for classification and characterization of wheat straw based on mechanical properties.* Industrial Crops and Products, 2018. **112**: p. 434-448.

[69]Yu, Z., et al., *Hybrid $ k $-nearest neighbor classifier.* IEEE transactions on cybernetics, 2015. **46**(6): p. 1263-1275.

[70]Syahfitri, K., S. Opim, and E. Syahril. *Sentiment analysis on Indonesian language text using support vector machine (svm) and k-nearest neighbor (kNN)(in Bahasa)*. in *Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA)*. 2015.

[71]Duneja, A. and T. Puyalnithi, *Enhancing classification accuracy of k-nearest neighbours algorithm using gain ratio.* Int. Res. J. Eng. Technol, 2017. **4**(9): p. 1385-1388.

[72]Mitani, Y. and Y. Hamamoto, *A local mean-based nonparametric classifier.* Pattern Recognition Letters, 2006. **27**(10): p. 1151-1159.

[73]Syaliman, K., E. Nababan, and O. Sitompul. *Improving the accuracy of k-nearest neighbor using local mean based and distance weight*. in *Journal of Physics: Conference Series*. 2018. IOP Publishing.

[74]Mirkes, E.M., J. Allohibi, and A. Gorban, *Fractional norms and quasinorms do not help to overcome the curse of dimensionality.* Entropy, 2020. **22**(10): p. 1105.

[75]Lee, T.-H., A. Ullah, and R. Wang, *Bootstrap aggregating and random forest.* Macroeconomic forecasting in the era of big data: Theory and practice, 2020: p. 389-429.

[76]Sage, A.J., U. Genschel, and D. Nettleton, *Tree aggregation for random forest class probability estimation.* Statistical Analysis and Data Mining: The ASA Data Science Journal, 2020. **13**(2): p. 134-150.

[77]Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Frontiers in neurorobotics, 2013. **7**: p. 21.

[78]Briscoe, E. and J. Feldman, *Conceptual complexity and the bias/variance tradeoff.* Cognition, 2011. **118**(1): p. 2-16.

[79]Greenwell, B.M., *pdp: An R package for constructing partial dependence plots.* R J., 2017. **9**(1): p. 421.

[80]Parr, T., J.D. Wilson, and J. Hamrick, *Nonparametric feature impact and importance.* arXiv preprint arXiv:2006.04750, 2020.

[81]Molnar, C., et al. *Relating the partial dependence plot and permutation feature importance to the data generating process*. in *World Conference on Explainable Artificial Intelligence*. 2023. Springer.

[82]Goldstein, A., et al., *Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation.* journal of Computational and Graphical Statistics, 2015. **24**(1): p. 44-65.

[83]Wright, R., *Interpreting black-box machine learning models using partial dependence and individual conditional expectation plots.* Exploring SAS® Enterprise Miner Special Collection, 2018. **1950**.

[84]Rozemberczki, B., et al., *The shapley value in machine learning.* arXiv preprint arXiv:2202.05594, 2022.

[85]Nóbrega, C. and L. Marinho. *Towards explaining recommendations through local surrogate models*. in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2019.

[86]Herbinger, J., et al. *Leveraging Model-Based Trees as Interpretable Surrogate Models for Model Distillation*. in *European Conference on Artificial Intelligence*. 2023. Springer.