scientific reports

Check for updates

OPEN

# Machine learning assisted exploration of the influential parameters on the PLGA nanoparticles

Sima Rezvantalab[1]✉, Sara Mihandoost[2]✉ & Masoumeh Rezaiee[1]

**Poly (lactic-co-glycolic acid) (PLGA)-based nanoparticles (NPs) are widely investigated as drug delivery systems. However, despite the numerous reviews and research papers discussing various physicochemical and technical properties that affect NP size and drug loading characteristics, predicting the influential features remains difficult. In the present study, we employed four different machine learning (ML) techniques to create ML models using effective parameters related to NP size, encapsulation efficiency (E.E.%), and drug loading (D.L.%). These parameters were extracted from the different literature. Least Absolute Shrinkage and Selection Operator was used to investigate the input parameters and identify the most influential features (descriptors). Initially, ML models were trained and validated using tenfold validation methods, and subsequently, next their performances were evaluated and compared in terms of absolute error, mean absolute, error and R-square. After comparing the performance of different ML models, we decided to use support vector regression for predicting the size and E.E.% and random forest for predicting the D.L.% of PLGA-based NPs. Furthermore, we investigated the interactions between these target variables using ML methods and found that size and E.E.% are interrelated, while D.L.% shows no significant relationship with the other targets. Among these variables, E.E.% was identified as the most influential parameter affecting the NPs' size. Additionally, we found that certain physicochemical properties of PLGA, including molecular weight (Mw) and the lactide-to-glycolide (LA/GA) ratio, are the most determining features for E.E.% and D.L.% of the final NPs, respectively.**

Nanomedicine, the science of designing and synthesizing medicines and imaging agents at the nanoscale, has garnered significant attention from the scientific community in recent decades. This field encompasses a wide array of materials that can serve as effective delivery platforms[1–3]. Since the beginning of the nanomedicine field, scientists from various backgrounds have tried to harness diverse knowledge to impel the advancements in the design of efficient nano-scaled pharmaceuticals. To boost the efficacy of nanomedicines and expedite their clinical application, a variety of strategies have been experimented with: co-loading of drugs and therapeutic agents (immunotherapy[4,5], phototherapy[6,7], imaging agents[8,9], etc.), surface modification, and functionalizing with moieties (targeting ligands[10], proteins[11,12], etc.). Experimentation and traditional trial-and-error practices are common, costly, and irregular in many of the strategies mentioned. ML-based methods, however, have recently gained prominence in the analysis and evaluation of drug delivery systems[13].

A variety of computational methods were utilized in this field, including quantum mechanics (at the electron, atomic, and molecular scales), molecular dynamics simulations (atoms and molecules moving at nanoscales), and mathematical and physiological pharmacokinetic/pharmacodynamic modeling (physicochemical behavior of drug delivery systems). Additionally, artificial intelligence (AI) based methods, including machine learning (ML) and deep learning have recently come into focus in this field, paving the way for data-driven nanomedicine. In comparison with other methods that are mostly based on theory, data-driven nanomedicine has gained a reputation for being based on experimental data as well as mathematical modeling and evaluations. Unlike other computational approaches, AI methods do not have a scale limit and can be applied to a wide range of applications. As an example, AI has been applied to the design of small drug molecules[14,15], as well as the control

[1]Chemical Engineering Department, Urmia University of Technology, Urmia 57166-419, Iran. [2]Present address: Electrical Engineering Department, Urmia University of Technology, Urmia 57166-419, Iran. ✉email: s.rezvantalab@uut.ac.ir; s.mihandoost@uut.ac.ir

of physicochemical properties of formulations (*e.g.* size, shape, drug loading efficiency, etc.) and even evaluation and optimization of (pre-)clinical data[16,17]. It is for this reason that these methods can help scientists cope with the complexity and ambiguity of practical DDSs, including polymer-based systems.

Its abilities to load and deliver a broad range of therapeutic and theranostic agents have made PLGA a preferred polymer-based DDS[18,19]. Due to its natural and non-toxic byproducts of PLGA degradation and its ability to conjugate various ligands, it has been approved as a DDS for a variety of formulations, including Decapeptyl® and several other microparticles based on PLGA, PGA, and PLA[20,21]. A wide range of therapeutic agents have been successfully delivered by PLGA-based (NPs), but factors that affect their physicochemical properties are complex and intertwined. For instance, Fredenberg et al.[22] discussed the complexity of the drug release mechanism and physicochemical properties of PLGA-based NPs. In addition to drug release mechanisms, various parameters can impact PLGA-based NPs' final performance and application. Previously, the determinant parameters and the complexity between them were portrayed as a Rubik's cube, where a small change in one brick (parameter) would affect the whole plane (other parameters and subsequent outcomes). NP size, drug loading and release are all influenced by the molecular weight (Mw) of the polymer, LA/GA ratio, solvent, and the synthesis method[23].

Since many parameters are ignored in traditional mathematical methods to avoid complexity, it appears impossible to establish a clear relationship between these effective parameters and their outcomes. For example, diffusion-driven models of drug-carrier interactions and diffusion-driven theoretical simulations of drug loading and release are limited by many assumptions. In this type of simulation, polymer segments are ignored due to the model complexity and treated as one polymer mass. These mathematical models also require approximate initial and boundary conditions. It is possible, however, to add each polymer's contribution to ML-based methods, and get a deeper understanding of their interactions. This type of method can provide more accurate and detailed information about material behavior, which can be used to predict the behavior of more complex systems based on the properties of the material. In this regard, using AI-assisted methods it is possible to design formulations with better control of properties and optimal outcomes[24–26]. For instance, multilayer perceptron artificial neural networks (ANNs) were used by Damiati et al.[27] to determine the sizes of (single or multiple) PLGA microparticles produced in microfluidic systems. PLGA solution concentration, microfluidic device geometry, and organic and aqueous flow rates have been used as input layers (effective factors) to train the ANN method. In their study, they reported highly accurate size predictions through microfluidics in the generation of PLGA-based microparticles. The same multilayer perceptron ANNs were used in a subsequent study to optimize indomethacin-loaded PLGA microparticle properties[28]. A successful application of the AI method was reported in the design of microparticles with satisfactory D.L.% and E.E.% (7.79 and 62.35%, respectively).

In this study, through the application of ML methods, we seek to advance our understanding of the critical factors involved in the synthesis and characteristics of PLGA-based NPs. Size, E.E.%, and D.L.% of the final NPs are influenced not only by polymer properties (such as LA/GA ratios, Mw, polyethylene glycol (PEG) layer) but also by numerous process related properties like synthesis methods (including nanoprecipitation, single or double emulsion, microfluidically), drug type, and of course solvents and presence of surfactants in the synthesis section. We collected all the essential information from the literature on PLGA-based NPs through data mining techniques. We evaluated the performance of various regression models on the data and also determined the order of importance of different features for every target. Additionally, we assessed the interactions between size, E.E.%, and D.L.% using ML techniques. At the end, we modeled data utilizing the effective features and the selected regression algorithm.

## Results and discussion

With the ML method, we aimed to recognize a relationship between multiple characteristics so that future investigations could be guided accordingly. In this regard, a database of PLGA-based NPs has been constructed that covers a broad range of information about the polymer chemistry and NPs' preparation method. Figure 1 outlines the main workflow of the current study.

The dataset is constructed with data from over a hundred research articles using keyword "PLGA NPs" which the detailed information was gathered all the reported factors. Eight features were identified as influential features in the final properties and performances of NPs. In this study, we selected the NPs' size, E.E.%, and D.L.% as significant parameters that can have a significant impact on formulation performance and therapeutic effect. Here are the definitions and measurements of E.E.% and D.L.% used in experimental studies:

$$E \cdot E \cdot \% = \frac{\text{Weight of entrapped drug in NPs}}{\text{Weight of total added drugs}} \times 100 \tag{1}$$

$$D \cdot L \cdot \% = \frac{\text{Weight of drug content in NPs}}{\text{Weight of NPs}} \times 100 \tag{2}$$

A number of features are polymer-related, for example, PLGA Mw, the presence of PEG in the formulation, and PEG Mw have the ability to influence the final properties of NPs. It worth mentioning that polymers' molecular weight has been used as an average of reported molecular weights. In addition, factors related to the preparation stage of NPs include the NPs preparation method (hereafter called the method), the solvents, and the presence of PVA (hereinafter called PVA). Data mining revealed that nanoprecipitation, single/double emulsion, and microfluidics methods are reported for the synthesis of NPs. Solvents used to dissolve PLGA polymers along with drugs to be loaded are another important aspect of NPs preparation that can affect final characteristics and performance. The literature review reports the use of numerous solvents in this process.

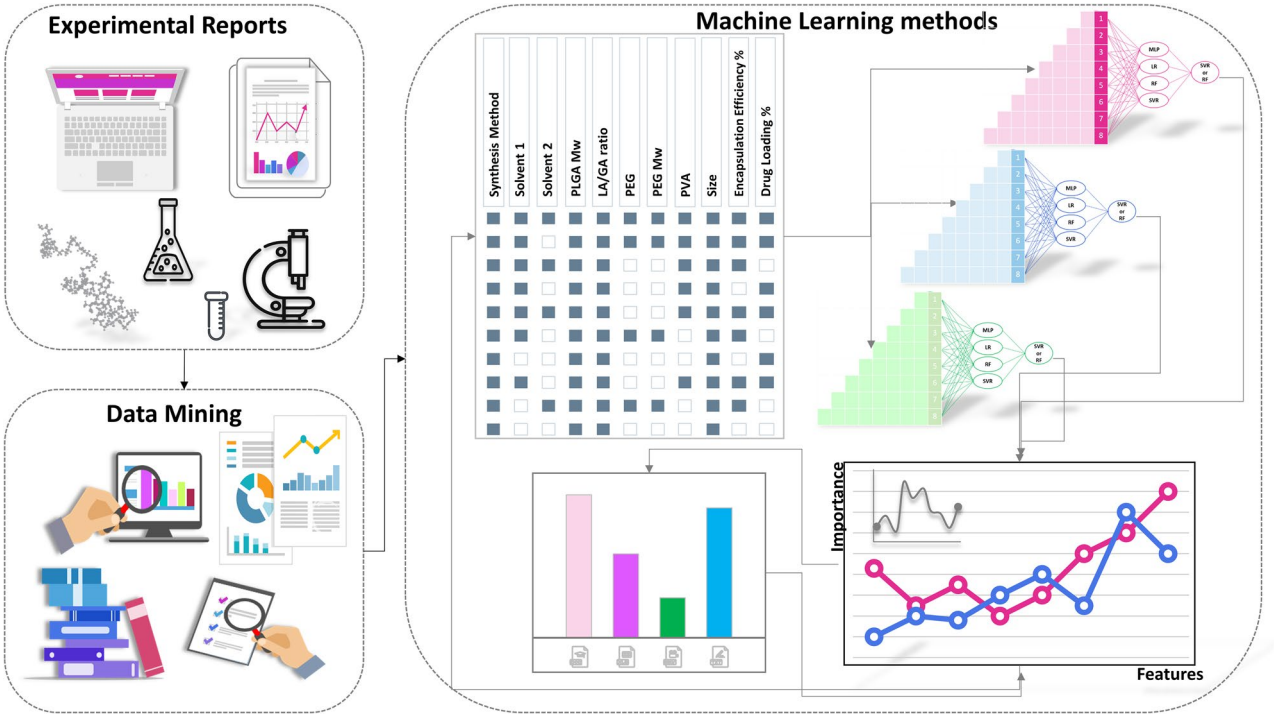


**Figure 1.** An overview of the current study's workflow. The reported data from experimental researches has been collected through a data mining stage. After a data cleaning step, ML-based algorithms have been employed to predict the effective parameters in the final characteristics of PLGA-based NPs.

## Feature and model selection

After compiling the dataset, our primary objective was to recognize and analyze the relationships within the gathered data. To achieve this, we employed a variety of well-established ML methods, including RF, SVR, LR, and MLP, and systematically applied them to the dataset. The hyperparameters of each ML model were carefully adjusted during the training and validation phases, utilizing a tenfold validation technique for each modelFurthermore, to evaluate the significance of different features (descriptors) and obtain a comprehensive overview of each descriptor's impact on the target and sort them by their weights, the feature set and target were subjected to the Lasso algorithm to assign a coefficient weight to each feature, as documented in Table 1. Subsequently, these coefficients were sorted and sequentially applied to the various ML algorithms, allowing us to compare their performance in terms of MSE, as illustrated in Fig. 2. The results obtained for size and E.E.% demonstrate that SVR consistently yields lower MSE across various feature numbers, outperforming other algorithms (Fig. 2a,b). In the case of D.L.%, although SVR performed reasonably well, LR exhibited the lowest MSE across different feature numbers, as shown in Fig. 2C.

Additionally, Fig. 2 illustrates the impact of different features on model performance. In the case of size (Fig. 2a), it is evident that when utilizing 7 features, which includes all features except LA/GA ratio, most ML models achieve the lowest level of MSE. Similarly, for E.E.%, all algorithms exhibit minimal error when considering 7 features, encompassing all features except Solvent 1. Regarding D.L%, Fig. 2c highlights that nearly all ML algorithms perform optimally with 7 features, which include all features except PEG Mw.

General view of results shows that the SVR model had the lowest MSE for size, E.E.%, while LR had best performance for D.L.%. The performance of SVR was with seven features *the* best among other ML models with

| | Size | E.E.% | D.L.% |
|---|---|---|---|
| Method | 0.23333 | 0.12025 | 0.13114 |
| Solvent 1 | 0.05756 | 0 | 0.24862 |
| Solvent 2 | 0.01376 | 0.0038 | 0.11565 |
| PLGA Mw | 0.13763 | 0.14582 | 0.27038 |
| LA/GA ratio | 0 | 0.05446 | 0.47369 |
| PEG | 0.27168 | 0.03584 | 0.0767 |
| PEG Mw | 0.12104 | 0.02146 | 0 |
| PVA | 0.22702 | 0.03216 | 0.43626 |

**Table 1.** Coefficient weight of each feature computed using Lasso algorithm.

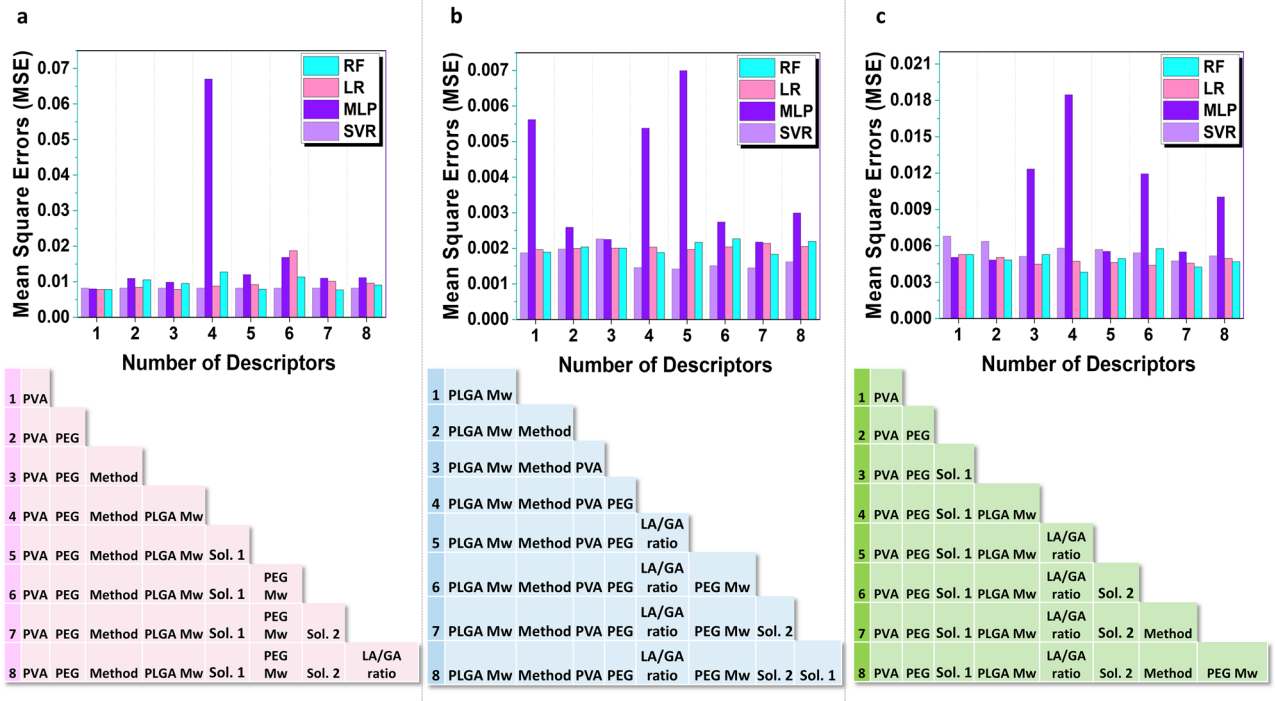


**Figure 2.** (**a**–**c**) Summary of the performance of various ML models for the prediction of size, E.E.%, and D.L.% using various features, respectively. Mean square errors (MSE) are considered the predictive performance measure in each case. A detailed description of the features and assigned numbers are included in the table under each diagram to avoid clumsiness in plots on the x-axis.

the lowest MSE ($0.0082 \pm 6.48 \times 10^{-6}$) and ($0.001696 \pm 0.00029$) respectively. In the case of D.L.%, LR had the best performance with six features, however; other methods such as MLP had the high MSE. Further scrutinizing showed that LR with seven features has the minimum MSE ($0.00486 \pm 0.00058$) whereas other algorithms also were in the low MSEs.

Furthermore, we conducted an additional experiment to validate the model selection. After determining the number of features for the models in the previous experiment, these selected features were applied to all ML algorithms to predict the respective targets, and Absolute Errors (AEs) were calculated. This experiment was carried out separately for size, E.E.%, and D.L.%. Figure 3 provides a summary of the results for each target, confirming the accuracy of the selected techniques in the previous experiment. For size and E.E%, the SVR method exhibited the lowest AE, as indicated by the smaller candle size and proximity to zero. In the case of D.L.%, LR demonstrated the lowest AE, as reflected in the candle size and proximity to zero.

## Assessing the interactions between different targets in the presence of the selected features

In this phase, we examined the interactions between different targets in the presence of the selected features. We utilized the LASSO algorithm to evaluate these interactions. For each target, we applied the LASSO algorithm to the 7th set of features (depicted in Fig. 2) along with the other two targets. An intriguing aspect of our approach was considering the influence of each target on the others. We hypothesized that size, E.E.%, and D.L.% exhibit interdependencies. For example, changes in E.E.% may impact the size of NPs, and vice versa. Consequently, when evaluating each target, we also took into account the other targets as input features.

Figure 4 provides an overview of the impact of all features on the size, E.E.%, and D.L.%, allowing for a comparison of their respective weights. Notably, E.E.% demonstrates the most significant influence on the size of the NPs, confirming the accuracy of our previous assumption, as depicted in Fig. 4a. Additionally, the presence of a PEG layer surrounding PLGA-based NPs is observed to have a crucial effect on their size, which has been well-documented in several studies[29–31]. For instance, previous research by Afshari et al.[32], found that PLGA NPs with PEG layers exhibited smaller sizes compared to bare PLGA NPs.

Additionally, the synthesis method as well as presence of PVA during NP production, can be very influential, according to the plot. Reports on the impact of PVA on the size of PLGA-based NPs are contradictory; some report a decrease[33,34] and others an increase[35,36]. Nevertheless, our findings are in good agreement with previous reports that confirm the NPs sizes are dependent on the presence of PVA. Furthermore, another effective descriptor as a synthesis method. Finally, it can be understood that Mw of both (PEG and PLGA) polymers are effective in determining the size of NPs. Based on a library of NPs produced by both PLGA and PEG polymers, both polymers have a significant impact on the final NP size[37,38]. Altogether, it is imperative that polymer physicochemical properties as well as synthesis methods be taken into consideration when designing and producing PLGA-based NPs.

In the subsequent step, we repeated this experiment using LASSO for the prediction of E.E.%. We observed that LASSO model identified the size among effective features on the E.E.%, according to Fig. 4b. Additionally,

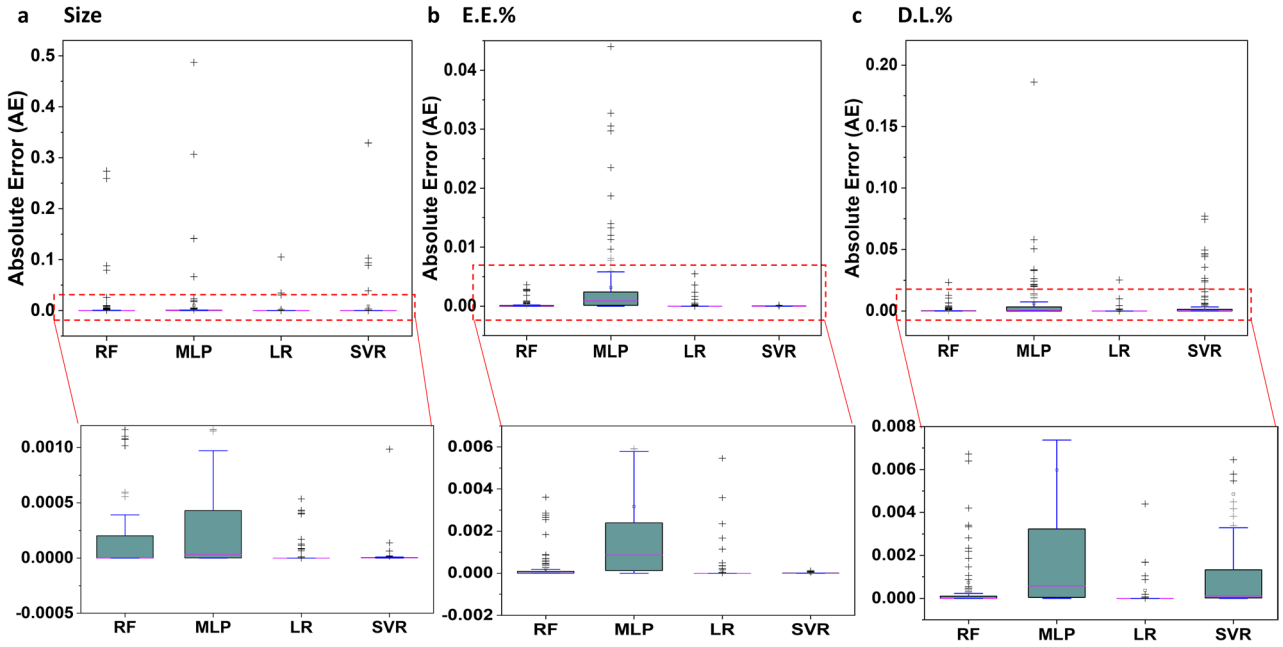


**Figure 3.** Validation of various ML models' performance based on absolute error for the selected sets of features. Subfigure (**a**) displays the absolute errors (AE) obtained with different ML models for the 7th set of feature descriptors used to predict size. Subfigure (**b**) demonstrates the AE obtained with different ML models for the 7th set of feature descriptors in predicting E.E.%. Lastly, Subfigure (**c**) exhibits the AE obtained with different ML models for the 7th set of feature descriptors in predicting D.L.%. To enhance the comparison of AE values, red lines are used to represent zoomed-in AE values.

the classifying model identified that the molecular weight of PLGA plays the biggest role in the importance of features in E.E.%. Our results are in good agreement with previous papers that have reported a correlation between size and E.E.%[39,40]. Furthermore, synthesis method is recognized as effective descriptor in the NPs E.E.%. Interestingly, it can be understood that the synthesis method acts as a double-edged sword that higher E.E.% gives rise into larger sizes. Therefore, feature researchers can design the NPs based on the trade-off between size and E.E.%. There is an interplay between size and E.E.%, as the third important effective descriptor on E.E.%. This observation is parallel with the previous finding and corroborates our hypothesis. Remarkably, analysis of feature weights form size and E.E.% reveals that D.L.% has no contribution in these targets and its weight in both cases is zero. This can be related to the different definitions of E.E.% and D.L.% (Eqs. 1 and 2, respectively). As mentioned it before, E.E.% refers to the proportion of encapsulated drug to the dissolved drug in a solvent. D.L.%, on the other hand, shows the weight of entrapped drugs in NPs over their total weight. Thus, this unexpected outcome was a direct result of the different definitions of E.E.% and D.L.%.

Similar evaluations have been conducted for D.L.% to identify its dependency on features. Firstly, testing the effective features on D.L.% led to the observation that there is no relationship between D.L.% and other targets (size and E.E.%), as expected (Fig. 4c). Additionally, LA/GA ratios can enormously affect D.L.% because, from a practical point of view, LA/GA ratio can alter the hydrophobicity and crystallinity of the polymers. In consequence, it may influence drug interactions with PLGA polymer chains, since it shifts the affinity of a drug for PLGA polymer. Obviously, D.L.% of NPs is also sensitive to the presence of PVA in formulations. In agreement with previous articles[41,42] and similar to E.E.%, D.L.% also depends on the presence and concentration of PVA. Again, we observe that polymer Mw acts as a double-edged sword in that the highest D.L.% is corresponding to the lowest E.E.%. Concludingly, Mw of polymer chains is one of the characteristics that researchers should take care of and choose according to their target in order to produce NPs without sacrificing one property for another.

Furthermore, to gain a better understanding of the relationship between the selected features and targets, we have prepared Fig. 5, which presents the correlation values between all features and targets. Correlation evaluates the linear relationship between two variables. A correlation value between 0 and 1 indicates a positive linear relationship, while a correlation value between -1 and 0 indicates a negative linear relationship. A correlation value of 0 implies that there is no linear relationship between the two variables. According to Fig. 5, we can investigate the linear relationship between the final feature set and the targets one by one. For example, the PEG and LA/GA ratio exhibit the highest correlation, with a value of 0.85. This indicates that when the PEG increases or decreases, the LA/GA ratio behaves in the same manner. On the other hand, the correlation between PEG Mw and E.E.% is − 0.37, indicating a linear relationship between PEG Mw and E.E.%, but in the opposite direction. Overall, these results suggest that linear correlations of inputs are useful in modeling the targets. However, it's important to note that correlation cannot capture non-linear relationships between variables. Therefore, further investigation is necessary to understand the contributions of individual inputs.

Furthermore, Table S1 outlines effects of each input feature on each target individually in terms of $r^2$.

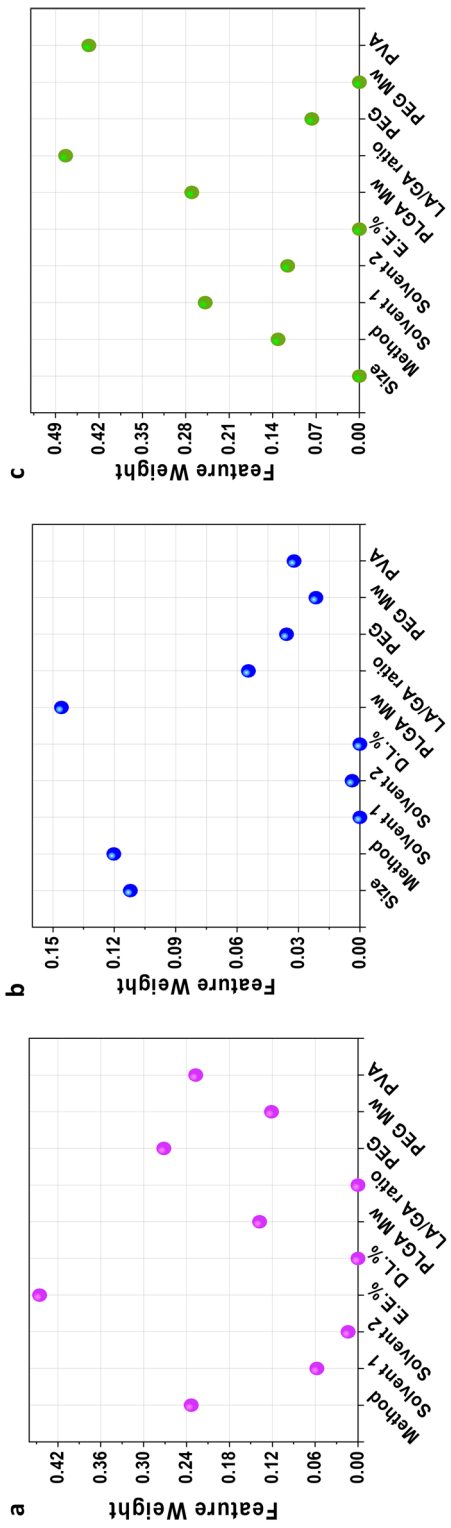


**Figure 4.** Effective features on the NPs' size. (**a**–**c**) Impact of various features on the size, E.E.%, and D.L.%, respectively.

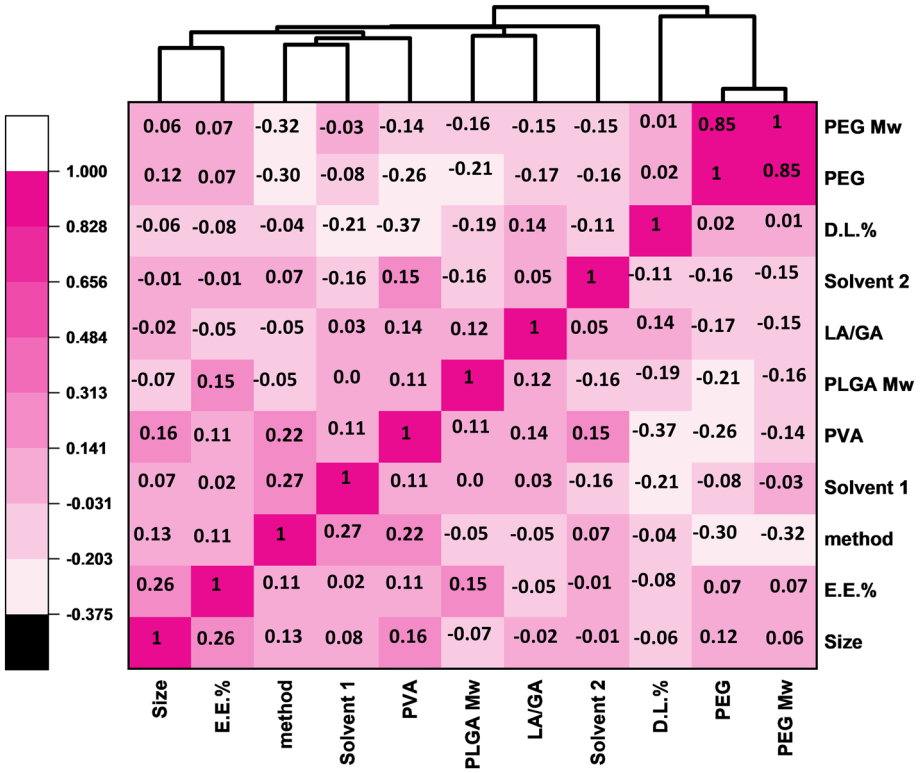


**Figure 5.** Correlation heatmap of all input features. A dark pink color indicates an absolute correlation (= 1) and a lighter pink color indicates a negative correlation.

## Model interpretation after descriptor selection

After the model and feature selection process, we focused on interpreting the models using the remaining *features*. In this stage, we removed the features that had zero weights in the previous section and evaluated the models using the non-zero features. Specifically, we excluded the features D.L.% and LA/GA ratio for size, as well as D.L.% and sol1 for E.E.% and size. Additionally, size, E.E.%, and PEG Mw were not considered for the DL case. To ensure consistency, all continuous variables were normalized to the range of 0 to 1. As mentioned earlier, we employed SVR with a radial basis function (RBF) kernel for predicting size and E.E.%, while LR was utilized for D.L.%. For SVR, we set the regularization parameter to 1 and the gamma parameter to 'scale'. In the case of LR, we used the binomial logistic regression model in logit mode. Figure 6 showcases the data for each target, with the selected features applied to the respective ML model. The scatter plots in the figure illustrate the training and testing results. It is evident that our proposed ML algorithm effectively learns from the training data and exhibits accurate predictions during the testing phase.

## Conclusion

In conclusion, with advancements in computer science research trends, we suggest a data-driven approach, to design and produce PLGA-based NPs to be used as efficient DDSs. For this purpose, ML techniques are used to exploit data from previous research articles and enhance the knowledge of algorithms on the relationships between multiple features. Our assessments revealed that the SVR model can be advantageous for the prediction of size and E.E.% of NPs with datasets containing 7 features, while LR showed the best performance with the same number of features for the prediction of D.L.%. Further evaluations revealed that size and E.E.% also interplay and are one of the most effectors among other features.

## Methods

### Data mining

The dataset for the current study was gathered from published articles about PLGA-based NPs. More than 200 studies were found through a search in google scholar using the keywords "PLGA NPs", "drug delivery", "drug loading", and "encapsulation efficiency". Afterward, reported information about size, E.E.%, D.L.%, presence of PEG and polyvinyl alcohol (PVA) in the formulation and synthesis, LA/GA ratio, Mw of PLGA and PEG, and solvent(s) used in the synthesis method was collected from papers. Table 2 outlines the data that were mined from papers with their numerical distribution. It should be mentioned, the high standard deviations (SD) of data in Table 2 indicates that the data points exhibit significant diversity around the mean. This diversity is important as it helps prevent ML techniques from becoming biased towards specific parts of the data. By having a larger SD, our dataset encompasses a wider range of values, allowing the ML algorithms to capture and learn from the

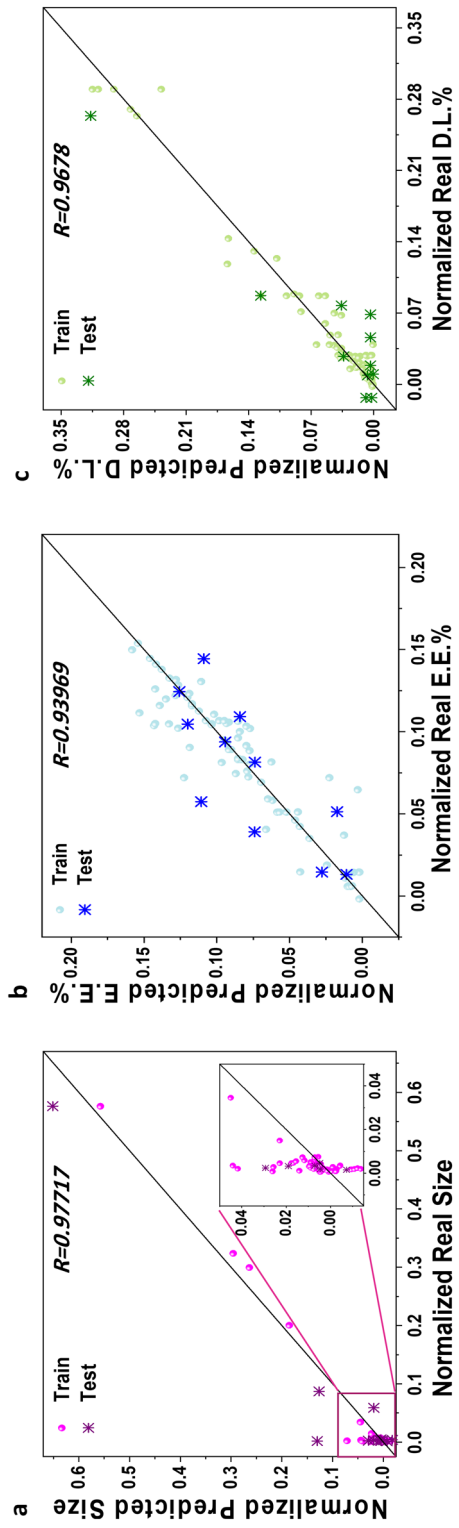


**Figure 6.** Relationship between the experimentally reported size, E.E.%, and D.L.%, predicted values.

| Input Features | | Target variables | |
|---|---|---|---|
| Name | Range | Name | Range |
| PLGA Mw (kDa) | 45.10 ± 38.7 | Size (nm) | 234.71 ± 180.2 |
| LA/GA ratio | 50/50, 65/35, 75/25, 80/20, 85/15, 95/5, 90/10 | E.E.% | 50.75 ± 26.5 |
| Presence of PVA | 0 or 1 | D.L.% | 16.95 ± 26.7 |
| Presence of PEG | 0 or 1 | | |
| PEG Mw (kDa) | 0.83 ± 1.79 | | |
| Solvent 1 | As Table 2 | | |
| Solvent 2 | As Table 2 | | |
| Synthesis Method | Nanoprecipitation, Double emulsion, Single emulsion, Microfluidics | | |

**Table 2.** The information collected from papers.

various patterns and variations present in the data. This ultimately enhances the robustness and generalizability of our ML models.

To construct a usable dataset for ML models training, machine-readable features were reported as *PLGA and PEG Mw (as the average of reported Mw range in the papers), LA/GA ratio (75/25, 50/50, 80/20, 65/35, 85/15, 95/5, and 90/10), synthesis method (nanoprecipitation, single emulsion, double emulsion, and microfluidics), presence of PVA and PEG (as a binary digit, one and zero for presence and absence, respectively, in the composition or synthesis method)*, and the list of solvents with their assigned codes, as in Table 3.

## ML-based models

In this study, we utilized various ML-based techniques, including multilayer perceptron (MLP), random forest (RF), logistic regression (LR), and support vector regression (SVR), to predict the size, E.E.%, and D.L.% of PLGA data.

To evaluate the performance of these ML models, we split the data into training and testing sets using a 70/30 ratio. During the training process, we fine-tuned the hyperparameters of each model using the tenfold cross-validation method, which helps in optimizing the model's performance and generalizability. It is worth mentioning that all the ML experiments were conducted using MATLAB 2022.A concise description of each model used is presented as follows:

*MLP*: In this study, we utilized an MLP network equipped with two hidden layers, one with 7 neurons and the other with 14 neurons, incorporating hyperbolic tangent sigmoid and logarithmic sigmoid transfer functions. The determination of the number of neurons in the hidden layers was a result of a trial-and-error process[43]. For training the neural netw *boost* ork in this study, we employed the Levenberg–Marquardt backpropagation method, conducted over 1000 epochs, while considering a minimum performance gradient of 1*e-7.

*RF*: is an ensemble learning method that excels in predicting continuous numerical values. The algorithm builds an ensemble of decision trees and combines their predictions to achieve accurate results. Let's denote the input features as $X = \{X_1.X_2.\cdots.X_n\}s$ where $n$ is the number of features, and the target variable as y. For each tree $T_i$ in the ensemble, the algorithm recursively partitions the data into subsets at each node, selecting the feature and split value that minimizes the mean squared error (MSE) of the target variable y. The final prediction $\widehat{y}$ for a given input $X$ is then obtained by averaging the predictions from all the individual trees in the forest:

$$\widehat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(X) \tag{3}$$

where N is the number of trees in the forest. This ensemble approach not only enhances the model's robustness and generalization but also allows for the calculation of feature importance scores, revealing the significance of

| Solvent | Code | Solvent | Code |
|---|---|---|---|
| Acetone | 1 | Dimethylsulfoxide | 8 |
| Dichloromethane | 2 | Tetrahydrofuran | 9 |
| Acetonitrile | 3 | Ethyl acetate | 10 |
| Chloroform | 4 | Hydrochloric acid | 11 |
| Ethanol | 5 | Dimethylformamide | 12 |
| Methanol | 6 | p-Dimethyaminobenzaldehyde | 13 |
| Sodium oleate | 7 | Methylene chloride | 14 |

**Table 3.** Reported solvents in the papers used for the preparation of PLGA NPs and their assigned readable codes for the ML models. Raw data will be available from the authors upon reasonable request. In this regards, readers can contact corresponding authors Sima Rezvantalab and/or Sara Mihandoost.

each feature in the prediction process. The power of RF lies in its ability to handle high-dimensional data, resist overfitting, and deliver reliable predictions in various regression applications[44].

*LR*: Logistic regression stands as a statistical technique employed to model the likelihood of a binary outcome, typically associated with events like success or failure. It hinges on the utilization of the logistic function, a transformative tool that maps the linear combination of input variables into bounded values within the range of 0 to 1. The formula that encapsulates logistic regression can be succinctly stated as:

$$P(Y=1)=\frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}} \tag{4}$$

In this equation, $P(Y=1)$ signifies the probability of the binary outcome assuming the value of 1, which in common scenarios corresponds to success. Key elements include $b_0$, the intercept, and $b_0.b_1.\cdots\cdot$ the coefficients aligned with the input variables $X_0.X_1.\cdots\cdot$ It's worth noting that $e$ represents the base of the natural logarithm. The logistic function's elegance lies in its ability to confine the output to a probability scale between 0 and 1, a fundamental characteristic that renders logistic regression particularly suitable for classification tasks[45].

*SVR*: is one of the most efficacious ML models for classification and regression. This method minimizes the structural risk by finding the most suitable decision function[46]. Briefly, using kernel functions, SVR can manage regression in high-dimensional datasets. Considering input and output vectors as X (from a $R^N$ input dataset) and Y such $\left\{(X.Y).X\in R^N.Y\in R\right\}$, respectively, SVR creates a linear function:

$$y=\omega^T\phi(X)+c \tag{5}$$

With $\omega$ and c are the weight matrix and the bias value, respectively. The feature point X corresponding to a data point is mapped from the initial space to the hyperspace, which may be defined as $\sum a_i k(X.X_i)=const$. Where $k(X.X_i)$ is the kernel function and $a_i$ are non-negative Lagrangian coefficients. Depending on the type of kernel function, the performance of SVR may be changed. In this paper, SVR has been trained with Gaussian or RBF kernels, both linear and polynomial, and the best mode that results from the practice of non-linear RBF kernel functions is considered. The best kernel function is RBF kernel functions, or $k(X_i.X)=\exp(\frac{-\|X_i-X\|^2}{2\sigma^2})$.

## Model evaluation

The absolute errors (AE) and the Mean square errors (MSE) were calculated via the following formulas to assess the performance of ML models:

$$AE=\left|\widehat{y}_i-y_i\right| \tag{6}$$

$$MSE=\frac{\sum_{i=1}^{n}\left|\widehat{y}_i-y_i\right|}{n} \tag{7}$$

As $\widehat{y}_i$ is the predicted target (size, E.E.%, or D.L.%) value; $y_i$ is the experimental value obtained from literatures. Total number of datasets in each target is shown with n.

## Feature Engineering: selection and reduction using LASSO

LASSO was first introduced by Robert Tibshirani, is a strong model for regularization and feature selection[47]. In this technique, the absolute value of the parameters has to be less than a fixed value (known as the upper bound) using a determined constraint in the method. In this regard, via a shrinking (regularization) process, the regression variables coefficients are penalized; some of them shrink to zero. Finally, non-zero coefficients are chosen to be part of the model.

This method aims to minimize the prediction error. In this method, λ determines the penalty, and with its larger values, the coefficients are dimensionality lowers down to zero. A higher λ value needs a greater number of coefficients that should be shrunk. However, its zero value gives rise to an Ordinary Least Squares (OLS) regression. LASSO predictions are very accurate since the bias doesn't change during the shrinking process. Additionally, by eliminating irrelevant variables, LASSO increases the interpretability of the model as well as decreases the overfitting. Regarding to the advantages, LASSO has been chosen for the feature selection task.

## Data availability

The data that support the findings of this study are gathered from published papers during last decade. Data will be available from the authors upon reasonable request.

Received: 3 November 2023; Accepted: 27 December 2023
Published online: 11 January 2024

## References

1. Ho, D., Wang, P. & Kee, T. Artificial intelligence in nanomedicine. *Nanoscale Horizons* **4**(2), 365–377 (2019).
2. Maleki, R., Rezvantalab, S. & Shahbazi, M.-A. Role of molecular simulation in the future of nanomedicine. *Future Med.* **16**, 2133–2136 (2021).
3. Lammers, T. & Ferrari, M. The success of nanomedicine. *Nano Today* **31**, 100853 (2020).
4. Koerner, J. *et al.* PLGA-particle vaccine carrying TLR3/RIG-I ligand Riboxxim synergizes with immune checkpoint blockade for effective anti-cancer immunotherapy. *Nat. Commun.* **12**(1), 1–16 (2021).
5. Lu, X. *et al.* Engineered PLGA microparticles for long-term, pulsatile release of STING agonist for cancer immunotherapy. *Sci. Transl. Med.* **12**(556), 6606 (2020).

6. Wang, H.-Q. *et al.* Indocyanine green accurately track the phototherapy based on W18O49@ PLGA nanostructure in vivo for visual treat. *Sci. Adv. Mater.* **11**(10), 1423–1432 (2019).
7. Khanom, J. I., Rezk, A., Park, C. H. & Kim, C. S. Near-infrared responsive synergistic chemo-phototherapy from surface-functionalized poly (ε-caprolactone)–poly (d, l-lactic-co-glycolic acid) composite nanofibers for postsurgical cancer treatment. *Biomacromolecules* **23**(9), 3582–3592 (2022).
8. Hashemi, M., Shamshiri, A., Saeedi, M., Tayebi, L. & Yazdian-Robati, R. Aptamer-conjugated PLGA nanoparticles for delivery and imaging of cancer therapeutic drugs. *Arch. Biochem. Biophys.* **691**, 108485 (2020).
9. Mariano, R. N., Alberti, D., Cutrin, J. C., Geninatti Crich, S. & Aime, S. Design of PLGA based nanoparticles for imaging guided applications. *Mol. Pharm.* **11**(11), 4100–4106 (2014).
10. Ahmad, E. *et al.* Ligand decorated biodegradable nanomedicine in the treatment of cancer. *Pharmacol. Res.* **167**, 105544 (2021).
11. Kim, J. *et al.* Functional-DNA-driven dynamic nanoconstructs for biomolecule capture and drug delivery. *Adv. Mater.* **30**(45), 1707351 (2018).
12. Sen Gupta, A. Bio-inspired nanomedicine strategies for artificial blood components. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* **9**(6), e1464 (2017).
13. Wang, W., Ye, Z., Gao, H. & Ouyang, D. Computational pharmaceutics-A new paradigm of drug delivery. *J. Control. Release* **338**, 119–136 (2021).
14. Díaz, Ó., Dalton, J. A. & Giraldo, J. Artificial intelligence: a novel approach for drug discovery. *Trends Pharmacol. Sci.* **40**(8), 550–551 (2019).
15. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **119**(18), 10520–10594 (2019).
16. Silva, A. *et al.* Output-driven feedback system control platform optimizes combinatorial therapy of tuberculosis using a macrophage cell culture model. *Proc. Natl. Acad. Sci.* **113**(15), E2172–E2179 (2016).
17. Lee, B.-Y. *et al.* Drug regimens identified and optimized by output-driven platform markedly reduce tuberculosis treatment time. *Nat. Commun.* **8**(1), 1–11 (2017).
18. Kapoor, D. N. *et al.* PLGA: a unique polymer for drug delivery. *Ther. Deliv.* **6**(1), 41–58 (2015).
19. Ghitman, J., Biru, E. I., Stan, R. & Iovu, H. Review of hybrid PLGA nanoparticles: Future of smart drug delivery and theranostics medicine. *Mater. Des.* **193**, 108805 (2020).
20. Blasi, P. Poly (lactic acid)/poly (lactic-co-glycolic acid)-based microparticles: An overview. *J. Pharm. Investig.* **49**(4), 337–346 (2019).
21. Rezvantalab, S. & Moraveji, M. K. Microfluidic assisted synthesis of PLGA drug delivery systems. *RSC Adv.* **9**(4), 2055–2072 (2019).
22. Fredenberg, S., Wahlgren, M., Reslow, M. & Axelsson, A. The mechanisms of drug release in poly (lactic-co-glycolic acid)-based drug delivery systems–a review. *Int. J. Pharm.* **415**(1–2), 34–52 (2011).
23. Rezvantalab, S. *et al.* PLGA-based nanoparticles in cancer treatment. *Front. Pharmacol.* **9**, 1260 (2018).
24. Imanparast, F., Faramarzi, M. A., Paknejad, M., Kobarfard, F., Amani, A., Doosti, M., Preparation, optimization, and characterization of simvastatin nanoparticles by electrospraying: An artificial neural networks study. *J. Appl. Polym. Sci. 133*(28): (2016).
25. Nguyen, C. N., Tran, B. N., Do, T. T., Nguyen, H. & Nguyen, T. N. D-optimal optimization and data-analysis comparison between a DoE software and artificial neural networks of a chitosan coating process onto PLGA nanoparticles for lung and cervical cancer treatment. *J. Pharm. Innov.* **14**(3), 206–220 (2019).
26. Baghaei, B. *et al.* Modeling and closed-loop control of particle size and initial burst of PLGA biodegradable nanoparticles for targeted drug delivery. *J. Appl. Polym. Sci.* **134**(33), 45145 (2017).
27. Damiati, S. A., Rossi, D., Joensson, H. N. & Damiati, S. Artificial intelligence application for rapid fabrication of size-tunable PLGA microparticles in microfluidics. *Sci. Rep.* **10**(1), 1–11 (2020).
28. Damiati, S. A., Damiati, S., Microfluidic Synthesis of Indomethacin-Loaded PLGA Microparticles Optimized by Machine Learning. *Front. Mol. Biosci.*, 595, (2021).
29. Sulaiman, T. N. S., Larasati, D., Nugroho, A. K. & Choiri, S. Assessment of the effect of PLGA Co-polymers and PEG on the formation and characteristics of PLGA-PEG-PLGA Co-block polymer using statistical approach. *Adv. Pharm. Bull.* **9**(3), 382 (2019).
30. Gan, M., Zhang, W., Wei, S. & Dang, H. The influence of mPEG-PCL and mPEG-PLGA on encapsulation efficiency and drug-loading of SN-38 NPs. *Artif. Cells, Nanomed., Biotechnol.* **45**(2), 389–397 (2017).
31. Song, Z. *et al.* Curcumin-loaded PLGA-PEG-PLGA triblock copolymeric micelles: Preparation, pharmacokinetics and distribution in vivo. *J. Colloid Interface Sci.* **354**(1), 116–123 (2011).
32. Afshari, M., Derakhshandeh, K. & Hosseinzadeh, L. Characterisation, cytotoxicity and apoptosis studies of methotrexate-loaded PLGA and PLGA-PEG nanoparticles. *J. Microencapsul.* **31**(3), 239–245 (2014).
33. Halayqa, M. & Domańska, U. PLGA biodegradable nanoparticles containing perphenazine or chlorpromazine hydrochloride: Effect of formulation and release. *Int. J. Mol. Sci.* **15**(12), 23909–23923 (2014).
34. Ranjan, A. P., Mukerjee, A., Helson, L. & Vishwanatha, J. K. Scale up, optimization and stability analysis of Curcumin C3 complex-loaded nanoparticles for cancer therapy. *J. Nanobiotechnol.* **10**, 1–18 (2012).
35. Budhian, A., Siegel, S. J. & Winey, K. I. Haloperidol-loaded PLGA nanoparticles: Systematic study of particle size and drug content. *Int. J. Pharm.* **336**(2), 367–375 (2007).
36. Ramirez, J. C. *et al.* Preparation of PDLLA and PLGA nanoparticles stabilized with PVA and a PVA-SDS mixture: Studies on particle size, degradation and drug release. *J. Drug Deliv. Sci. Technol.* **60**, 101907 (2020).
37. Valencia, P. M. *et al.* Microfluidic platform for combinatorial synthesis and optimization of targeted nanoparticles for cancer therapy. *ACS Nano* **7**(12), 10671–10680 (2013).
38. Bertrand, N. *et al.* Mechanistic understanding of in vivo protein corona formation on polymeric nanoparticles and impact on pharmacokinetics. *Nat. Commun.* **8**(1), 777 (2017).
39. Qi, F. *et al.* Preparation of uniform-sized exenatide-loaded PLGA microspheres as long-effective release system with high encapsulation efficiency and bio-stability. *Colloids Surf., B* **112**, 492–498 (2013).
40. Shubhra, Q. T. *et al.* Co-encapsulation of human serum albumin and superparamagnetic iron oxide in PLGA nanoparticles: Part II. Effect of process variables on protein model drug encapsulation efficiency. *J. Microencapsul.* **31**(2), 156–165 (2014).
41. Behnke, M. *et al.* Optimized encapsulation of the FLAP/PGES-1 inhibitor BRP-187 in PVA-stabilized plga nanoparticles using microfluidics. *Polymers* **12**(11), 2751 (2020).
42. Park, H. *et al.* Effect of stabilizers on encapsulation efficiency and release behavior of exenatide-loaded PLGA microsphere prepared by the W/O/W solvent evaporation method. *Pharmaceutics* **11**(12), 627 (2019).
43. Gholizadeh, M., Jamei, M., Ahmadianfar, I. & Pourrajab, R. Prediction of nanofluids viscosity using random forest (RF) approach. *Chemom. Intell. Lab. Syst.* **201**, 104010 (2020).
44. Liu, L. *et al.* Cytotoxicity of phytosynthesized silver nanoparticles: A meta-analysis by machine learning algorithms. *Sustain. Chem. Pharm.* **21**, 100425 (2021).
45. LaValley, M. P. Logistic regression. *Circulation* **117**(18), 2395–2399 (2008).
46. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**(5), 1207–1245 (2000).
47. Tibshirani, R. The lasso method for variable selection in the cox modeL. *Statist. Med.* **16**(4), 385–395 (1997).

## Author contributions

The manuscript was written through the contributions of all authors. All authors have given approval for the final version of the manuscript. S.R. conceptualized and wrote the manuscript, and S.M. performed all the modeling and computational sections. S.R. and S.M. interpreted the results. M.R. gathered the data and contributed to the data mining stage.

## Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-50876-w.

**Correspondence** and requests for materials should be addressed to S.R. or S.M.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.