PHARMACIA

Pharmacia 71: 1–7
DOI 10.3897/pharmacia.71.e122772

Research Article

A data-driven approach to predict the *in vitro* dissolution time of sustained-release tablets using raw material databases and machine learning algorithms

M. Bharathi[1], Raju Kamaraj[1], S. Murugaanandam[2], Kota Navyaja[1], T. Sudheer Kumar[1]

1 *Department of Pharmaceutical Regulatory Affairs, SRM College of Pharmacy, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu -603203, India*

2 *Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu -603203, India*

Corresponding author: Raju Kamaraj (kamarajr@srmist.edu.in)

Received 11 March 2024 ♦ **Accepted** 8 August 2024 ♦ **Published** 26 August 2024

Citation: Bharathi M, Kamaraj R, Murugaanandam S, Navyaja K, Kumar TS (2024) A data-driven approach to predict the *in vitro* dissolution time of sustained-release tablets using raw material databases and machine learning algorithms. Pharmacia 71: 1–7. https://doi.org/10.3897/pharmacia.71.e122772

Abstract

Tablets are the most typical dosage forms of pharmaceutical inventions. Sustained-release (SR) tablet formulations are designed to release the drug gradually in the bloodstream and often require less frequent dosing. Current strategies to optimize sustained-release tablet dissolution time still rely on the traditional approach, which is time-consuming and expensive. In the present context, we have demonstrated alternate machine learning and deep learning models through the TPOT AutoML platform. Six machine learning (ML) models were compared to improve the methodology for dissolution time prediction, particularly the decision tree regressor (DTR), gradient boost regressor (GBR), random forest regressor (RFR), extra tree regressor (ETR), XGBoost regressor (XGBR), and deep learning (DL). The obtained results indicated that machine learning methods are convincing in speculating the dissolution time, especially the random forest regressor, but upon hypertuning of the deep neural network, the deep learning model with a 10-fold cross-validation scheme demonstrated superior predictive performance with an NRMSE of 8% and an R^2 of 0.92. The major essentials affecting the dissolution time of SR tablets were explained using the SHAP method.

Keywords

dataset development, deep learning, machine learning model, SHAP method, TPOT autoML

Introduction

Pharmaceutical design needs to be tailored for all novel drug entities due to the variability in these relationships (Hayashi et al. 2023). The primary goal of many pharmaceutical drugs is to attain a consistent blood level that is both helpful for therapeutic purposes and safe over a long period of time (Mishra et al. 2019). Therefore, sustained-release (SR) tablet dosage forms are intended to achieve an extended therapeutic impact. SR dosage forms aim to enhance drug efficacy by prolonging therapeutic effects, allowing for less frequent dosing, and utilizing lower

doses (Bhagat et al. 2023). They are also considered a wider choice for drugs with quick life spans, which typically need regular administration to sustain their therapeutic levels. Overall, a lower dosage and less frequent dosing can contribute to fewer complications from the drugs. This is particularly beneficial in the treatment of chronic diseases, where long-term adherence to medication is crucial (Mishra et al. 2019). There has been a notable increase in SR drug delivery systems over the last two to three decades due to various factors like the high cost of raising the latest drug candidates, the running out of current patents, and the development of newly discovered polymers that enable controlled drug release (Bhagat et al. 2023).

The choice of polymer material, including hydrophilic and hydrophobic substances or biodegradable materials, shows the versatility of formulations. Studies of *in vitro* dissolution can be used to determine the medication release rate (Santoshrao et al. 2014). A formulation can possess many elements, like polymers, lubricants, solubilizers, binders, and fillers. The choice of formulation components depends on the specific dosage form and the requirements of the formulation procedure. Each element and its specific quantity directly impact the overall critical quality attributes (CQA) of a dosage form (Szlek et al. 2022).

Current strategies to optimize the dissolution time of SR tablets still rely on conventional trial-and-error methods, which are considered time-consuming and expensive. There may be a need for more efficient and streamlined approaches to enhance the optimization process (Yanga et al. 2019). One way to disrupt the conventional approach is by executing machine learning (ML) and deep learning (DL) models. ML has indeed become a transformative force in various research domains, including pharmaceuticals. ML models can make data-driven predictions with datasets of preliminary data.

ML methods, such as optimization, can be employed to find optimal combinations of ingredients and process parameters that produce the required properties, such as drug release profiles. This can lead to significant cost savings in terms of materials, resources, and time. ML models can contribute to maintaining product consistency by identifying key factors that influence the quality of formulations (Sustersic et al. 2023). DL is indeed a subset of ML that has gained significant attention and success in various domains of artificial intelligence. DL includes training artificial neural networks with several layers (DNNs) to learn and make predictions or choices.

This method has proven to be quite successful in a variety of applications, including natural language processing, image and speech recognition, and many others. Therefore, DL has indeed demonstrated remarkable success in various domains, often outperforming traditional ML methods in predictive performance. *In vitro*, prediction of a pharmaceutical formulation (Yanga et al. 2019), prediction of adverse drug reactions in drug discovery (Mohsen et al. 2021), prediction of pharmacological drug properties (Aliper et al. 2016), and DL-based dose prediction (Gronberg et al. 2023)—as many such works relaying on DL models were published in the last 5 years. The prediction of dissolution release by artificial neural network (ANN) and regression methods, respectively (Wang et al. 2022), prediction of HPMC polymer matrix particle size distribution (PSD) by using ANN, support vector machine (SVM), and ETR (Galatta et al. 2021), surface area and volume of the tablet affect the release profile of the tablet can be predicted by ANN (Mazur et al. 2023), and physiochemical and powder characteristics of API are predicted by four-layered ANN (Takayama et al. 2017). Multivariant tools and ANN methods have been seen to execute well in the context of understating the underlying relationship between CQAs and process parameters.

Nevertheless, ML and DL modeling methods can be used to assess the dissolution time of SR tablets with the above-mentioned inferences. However, with more comparisons of DL with ML, more approaches are looking forward to anticipating the effective formulation. In the present context, we developed an optimal method for the evaluation of the dissolution time of SR tablets based on ML models, namely DTR, GBR, XGBoost regressor, RFR, ETR, and DL. All these models have demonstrated successful applications in the fields of pharmaceutical formulation development, development processes, and destructive analytical tests. Therefore, by leveraging data-driven approaches, these methods contribute to more efficient and effective pharmaceutical development and manufacturing processes. Nevertheless, it is essential to remember that the successful implementation of these techniques requires careful consideration of data quality, model interpretability, and regulatory compliance in the pharmaceutical industry (Loua et al. 2021). This paper showcases a wide range of ML and DL methods for accurately forecasting the dissolution time of SR tablets.

Methodology

Pharmaceutical data description

An existing literature-based data model (Hana et al. 2018) was chosen for development and organization. Data was categorized to incorporate only data records with emphasis on SR tablets and their components like tablet hardness, thickness, friability, drug content, dissolution time, and dissolution release profile. To extend our database, we have done a literature review from the Scopus, Web of Sciences, and Google Scholar databases. The search engine operates with SR tablet formulations as its foundation: “HPMC” or “hydroxypropyl methylcellulose” or “hydroxypropylmethylcellulose” or “hydroxypropylmethyl cellulose” or “Hypromellose tablets (Yanga et al. 2019).

A total of 210 articles were retrieved through database search; from these, only 152 research articles have been picked for feature data extraction. Later, upon hand search, 80 articles did not match all the inclusion norms and were eventually removed from the study. After sorting 72 articles, a total of 1215 formulations were retrieved. The formulation data includes the name API and other excipients; process details were documented in the data-

set. The final dataset includes the following parameters for each formulation: API name, dose, excipient name, dose (each excipient as a separate column), hardness, friability, thickness, drug content, and cumulative drug release (Momeni et al. 2023).

Workflow

Three stages make up the process of developing an ML model: preprocessing the data, modeling, and model interpretation. These stages come after ML's best execution.

Preprocessing of data

After the completion of data collection, the data needs to be processed before building the predictive models to ensure the robustness and effectiveness of the ML models. A few normally employed methods, including data cleansing, dimension reduction, imbalanced data solutions, and data splitting strategies, are required to analyze the data. Data cleansing is used for observations of missing datasets, and it is employed by replacing the data points with median/mean values. Besides, there are few constraints on replacing the missing values; therefore, a decrease in data size might have an impact on the exactness of the model. The dimensionality reduction method is used to eliminate the dataset's least significant features, which will lower the overfitting issues and reduce the complexity of the model.

Different approaches to dimensionality reduction, including principal component analysis (PCA), high correlation filtering, and random forest feature selection, came out mostly in data processing. Imbalanced data solutions mostly discuss the uneven scattering of different database classes. Usage of an unbalanced dataset by a prediction model leads to non-significant performance. Data splitting is one more key step in data processing. In this method, the entire dataset is randomized and split into three subcategories: training, validation, and testing. The training set will train the models. Validation is for tuning the hyperparameters and stopping overfitting. A testing set is used to determine the prediction ability of unknown data. The desired ratio for these three categories is 70:20:10; moreover, the ratios rely on the data size. Consequently, data preprocessing and splitting strategies are essential steps before performing the task (Jiang et al. 2022).

Modelling

ML modeling tasks are framed to cover various techniques, including classification, regression trees, neural networks, and potentially many other algorithms. These models are trained using prepared databases. The performance of different models is evaluated using an error metric. Common error metrics include accuracy, precision, recall, the F1-score for classification tasks, and the normalized root mean squared error, R^2 score for regression tasks. Keeping track of different modeling methods and exploring various features can be challenging and computationally expensive. Therefore, automated machine learning (AutoML) is utilized. AutoML approaches frequently use ensemble learning strategies, which combine several model types to produce predictions that may be more reliable. In this case, the K-fold cross-validation technique was employed by TPOT AUTOML to generate a definitive production model by selecting features based on a predefined threshold. A distinct training-testing pair, consisting of 568 records randomly selected for training, 244 records for validation, and 348 records for testing, makes up each fold.

Model training

After the completion of the ML modeling process, it is essential to assess the predictive accomplishments of ML models to understand how well they generalize to new, unseen data. ML models are known to be overfit. Overfitting happens when the model learns not only the underlying patterns in training data but also the noise and random fluctuations that accompany the data. To avoid overfitting and maintain a stable model. The choice of features offers insights into which features had the biggest impact on the model's assumptions, revealing the black-box nature of ML models. Also, a subset of the data was excluded for evaluating the model using the K-fold approach. In this study, models were trained and verified via a five-fold cross-validation scheme, then split for feature selection by a Python script. The training and validation procedures were repeated 50 times to ensure complete coverage of the input database and get the best model. The final model was trained using a 10-fold cross-validation approach after the final input feature vector was chosen. Root mean square error (RMSE), normalized root mean squared error (NRMSE), and coefficient of determination (R^2) can be used to measure the correctness of the models. Six algorithms drawn from the TPOT AUTOML platform were utilized for feature selection and final model development: DTR, GBR, XGBoost, RFR, ETR, and DL.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(pred_i - obs_i\right)^2}{n}}$$

$$NRMS = \frac{RMSE}{obs_{max} - obs_{min}}(100\%)$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}\left(pred_i - obs\right)^2}{\sum_{i=1}^{n}\left(obs_i - obs\right)^2}$$

where "obs_i, $pred_i$" are the practical and expected values, "i" is the data record number, "n" is the total number of records, "obs_{max}" is the highest experimental value, "obs_{min}" is the least observed value, "R^2" is the coefficient of determination, "SS_{res}" is the sum of squares of the residual errors, "SS_{tot}" is the total sum of the errors, and "obs" is the arithmetical mean of observed values (Szlek et al. 2022).

Machine learning models

The accuracy of ML results cannot be improved simply by fitting data into models. As the data gets large and complex, better data handling techniques are essential to handling it.

Model interpretation

As ML models are inherently black boxes, efforts are undertaken to teach their prediction technique. In our work plan (Fig. 1), we used Lundberg et al.'s SHAPLEY additive explanation (SHAP) approach to describe the universal relationship between input and output variables. The SHAP method is indeed a mostly used path derived from cooperative game theory, and it has found applications in various domains, including the pharmaceutical domain. The notion of Shapley values is used to evaluate the beneficence of each participant (or individual) towards the overall team effort or outcome. In machine learning, the concept of Shapley values has been adapted to explain the contribution of each feature in a predictive model. SHAP values provide a way to allocate the model's prediction to each feature fairly and consistently. The mathematical formula is displayed in the following equation:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\}) - val(S))$$

while "S" is a subset of the features employed in the model, "x" is the vector of feature values to be elucidated, and "p" is the number of features. "$Val_x(S)$" is the prediction for feature values in set "S" that are marginalized over features excluded in set "S" (Fadel et al. 2022).

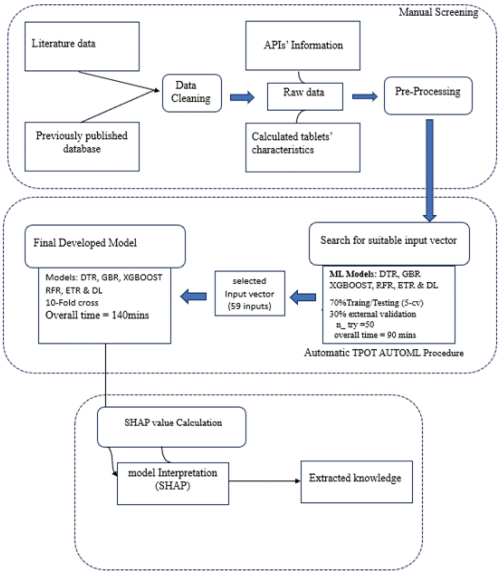

Figure 1. Schematic representation of applied workflow.

Results

Database

The pre-processed database contained 72 wet granulation SR formulations (new data entries), including 24 unique APIs. 50 variables encrypting composition (excipients were topologically encoded), 6 variables encrypting formulation dimensions such as thickness [mm], hardness [N], friability [%], drug content [%], dissolution studies, and dissolution time [Hrs]. Descriptive statistics: when topological encrypting was used, Table 1 demonstrates that the variables were not distributed normally; instead, the proportions of formulations were notably positively skewed (right-skewed distribution). However, the database was split using a 10-fold cross-validation approach that was balanced to ensure that the input variables were classified fairly throughout the splits.

Choosing features and creating the final model

Choosing features and creating the final model were accomplished by the TPOT AUTOML method. The AUTOML dimensions are given in Table 2. This table exhibits the accuracy of the developed models. The values of NRMSE, RMSE, R^2, MAE, and MSE are further analyzed to check the accuracy and precision of the model output.

As expected, ML techniques, especially RFR, have proved to be the best pipeline in the TPOT AUTOML analysis for the curated dataset, but it is pretty clear that the DL model, when hyper-tuned, performed on par with its ML counterparts with great R^2 values and NRMSE%. Following the preliminary evaluation of the DL models when trained using a five-fold CV scheme, when the DL model was trained with three hidden layers of 100 neurons each and 35 input neurons and one output neuron with epoch values of 2200 combined with a 10-fold cross-validation scheme, the model accuracy jumped and showed significant improvement in NRMSE and R^2 value.

Feature selection of input variables based on scaled importance

The input variables were categorized into two main groups: composition and manufacturing parameters. Except for those in the composition group, features from all categories that fell below the variable importance threshold were eliminated. Consequently, the final input vector contained 58 inputs.

According to the scaled importance of input variables, it shows a higher number of polymers. It is evident that a higher concentration of polymer has a high influence on the predicted values.As stated in Table 3, lubricants and solubilizers (PEG) are the least essential components of the tablets, which could be related to a positive skew in the variable distribution.

Model performance

The SHAP summary plots were performed to figure out the effect of two categories of inputs, namely formulation setup and manufacturing specifications, on output. Higher dissolution times are predicted where higher concentrations of polymers like Polyox WSR, HPMC, and carbopol occur. Regarding diluents (lactose, MCC) with lactose, a reverse correspondence is seen where a higher amount of lactose produces a low dissolution time. Re-

Table 1. Descriptive statistics of the dataset.

Feature	Count	Min	Mean	Std	25%	50%	75%	Max
Thickness (mm)	1160	4.01	15.95	0.3	2.53	3.23	4.1	445
Hardness(N)	1160	7.79	18.61	0.6	4.5	5.53	6.6	169
Friability (%)	1160	0.46	1.82	0	0.22	0.36	0.53	44
Drug content (%)	1160	74.69	41.47	1.74	50	98.34	99.42	106.56
Dissolution Release (%)	1160	106.22	408.32	9.56	86.63	95.99	98.89	9920
Dissolution Studies Time (Hrs)	1160	13.83	5.78	2	10	12	17	24
Solubilizer Mannitol [%]	1160	0.15	1.47	0	0	0	0	30
Solubilizer PEG [%]	1160	0.03	0.36	0	0	0	0	5
Polymer CMC [%]	1160	0.04	0.9	0	0	0	0	22.86
Polymer CMC Sodium [%]	1160	0.12	3.26	0	0	0	0	75
Polymer Carbopol [%]	1160	1.49	6.74	0	0	0	0	64.52
Polymer Carrageenan [%]	1160	0.02	0.37	0	0	0	0	10
Polymer Chitosan [%]	1160	0.21	1.87	0	0	0	0	27.84
Polymer Ethyl cellulose [%]	1160	1.49	6.82	0	0	0	0	59.69
Polymer Eudragit [%]	1160	1.94	7.76	0	0	0	0	56.93
Polymer Guar Gum [%]	1160	1.18	5.55	0	0	0	0	57.14
Polymer HPC [%]	1160	0.05	0.99	0	0	0	0	20
Polymer HPC 75–100 [%]	1160	0.08	1.13	0	0	0	0	22.86
Polymer HPC(M) [%]	1160	0.05	0.92	0	0	0	0	23.08
Polymer HPCK100M [%]	1160	0.16	3.26	0	0	0	0	75
Polymer HPMC [%]	1160	10.84	14.81	0	0	0	19.55	75
Polymer Karaya Gum [%]	1160	0.21	2.04	0	0	0	0	32.43
Polymer Metalose60SH50 [%]	1160	0.06	1.08	0	0	0	0	25
Polymer Methocel [%]	1160	0.07	0.96	0	0	0	0	15.67
Polymer Metolose [%]	1160	0.11	1.36	0	0	0	0	21.21
Polymer PEG [%]	1160	0.28	3.31	0	0	0	0	57.14
Polymer PVP [%]	1160	0.08	1.76	0	0	0	0	48
Polymer Polyox WSR [%]	1160	0.74	5.66	0	0	0	0	67.16
Polymer Polyoxyethylene [%]	1160	0.09	1.61	0	0	0	0	39.67
Polymer Sodium Alginate [%]	1160	0.8	6.03	0	0	0	0	72.99
Polymer Tara Gum [%]	1160	0.15	2.37	0	0	0	0	50
Polymer Tragacanth Gum [%]	1160	0.11	2.26	0	0	0	0	57.14
Polymer Xanthane Gum [%]	1160	1.45	6.84	0	0	0	0	72.99
Diluent Avicel [%]	1160	2.01	7.55	0	0	0	0	70.6
Diluent Dicalcium Phosphate [%]	1160	2.87	11.38	0	0	0	0	67.19
Diluent Lactose [%]	1160	12.06	21.94	0	0	0	13.51	86.7
Diluent MCC [%]	1160	15.43	23.24	0	0	0	28.15	93.33
Diluent Starch [%]	1160	0	0.11	0	0	0	0	3.85
Binder Maltodextrin [%]	1160	0.37	3.06	0	0	0	0	65.56
Binder Povidone [%]	1160	0.09	0.58	0	0	0	0	4.53
Binder Starch [%]	1160	0.71	4.97	0	0	0	0	56.75
Lubricant Aerosil [%]	1160	0.13	0.41	0	0	0	0	2.06
Lubricant Avicel [%]	1160	0.1	0.79	0	0	0	0	8.26
Lubricant Calcium Stearate [%]	1160	0.04	0.5	0	0	0	0	7.14
Lubricant Colloidal Silicon dioxide [%]	1160	0.06	0.81	0	0	0	0	14.06
Lubricant Magnesium Stearate [%]	1160	1.06	1.18	0	0	0.86	1.58	6
Lubricant Talc [%]	1160	0.26	0.73	0	0	0	0	4.73
Glidant Aerosil [%]	1160	0.09	0.62	0	0	0	0	8
Glidant Colloidal Silicon dioxide [%]	1160	0	0.03	0	0	0	0	0.49
Glidant HPMC [%]	1160	0.02	0.24	0	0	0	0	3.97
Glidant Magnesium Stearate [%]	1160	0.44	0.74	0	0	0	1	4.3
Glidant Talc [%]	1160	0.97	1.6	0	0	0	1.46	10

garding the lubricants (magnesium stearate, Aerosil, and Avicel), two different effects can be seen at high concentrations of Aerosil, leading to greater dissolution time, which may be due to its hydrophilic nature. The lipophilic nature of the Mg stearate decreases the dissolution times when its concentration is higher. For binders like povidone and maltodextrin, it is observed that higher dissolution times are associated with a higher concentration of povidone. Glidants (Talc, Aerosil): a decreased amount of Aerosil leads to increased dissolution time.

Table 2. Robustness of the TPOT AUTOML-developed models and the DNN model.

ML Techniques	RMSE (s)	NRMSE (%)	R^2	MAE	MSE
DTR	2.53	13	0.81	0.93	6.93
GBR	2.98	15	0.73	2.17	8.89
XGBOOST	2.15	11	0.86	1.18	4.68
RFR	2.05	10	0.87	1.1	4.19
ETR	2.24	11.0	0.58	1.2	5.0
DL (5-F-CV) DL (10-F-CV)	2.5	13	0.82	1.3	6.28
	1.61	8	0.92	0.96	2.59

Table 3. Selected input variables for the best predictive ML models.

Feature	Feature type	Scaled feature importance
HPC(M) [%]	Polymer composition	1
Talc [%]	Glidant composition	0.508057894
Polyoxyethylene [%]	Polymer composition	0.501707214
Sodium CMC [%]	Polymer composition	0.464332229
Dissolution Studies [%]	Manufacturing parameter	0.460079275
Dicalcium Phosphate [%]	Diluent composition	0.38616999
Colloidal Silicon dioxide [%]	Glidant composition	0.351912377
Solubilizer PEG [%]	Solubilizer composition	0.320415496
Carrageenan [%]	Polymer composition	0.296401411
HPMC [%]	Glidant composition	0.271820314
CMC [%]	Polymer composition	0.232133437
Povidone [%]	Binder composition	0.221659485
Magnesium Stearate [%]	Glidant composition	0.204297449
Thickness	Manufacturing parameter	0.110086855
Hardness	Manufacturing parameter	0.102620188
Maltodextrin [%]	Binder composition	0.100213431
Guar Gum [%]	Polymer, composition	0.091423699
Carbopol [%]	Polymer, composition	0.071688362
Colloidal Silicon dioxide [%]	Lubricant, composition	0.068848904

Discussion

Solid oral dosage forms occupy a major part in the field of pharmaceuticals, and their efficacy is highly dependent on the absorption of API in the human body. The major aspect that impacts the API is its dissolution behavior, and that is essential for ensuring the bioavailability and therapeutic effectiveness of the API. However, the dissolution profile is mostly influenced by the physical characteristics of the materials, like solubility and filling material employed in the development procedure, and the process variables, like compression force. Therefore, it often involves a careful balance of formulation design, choice of materials, and control of manufacturing processes.

Regulatory agencies often set specifications for dissolution profiles to ensure the consistency and efficacy of the product. The traditional method for dissolution involves testing a small number of tablets, which may be laborious, prolonged, and overpriced. Additionally, it would not provide a complete representation of the overall dissolution behavior. As a result, there is interest in developing surrogate or alternate methods that can provide insights into dissolution attributes without the need for extensive testing (Fink et al. 2023). On account of this, an alternative method was employed: ML models are perhaps powerful function approximators and can memorize and forecast composite systems from huge amounts of input.

Lately, ML has been productively implemented in pharmaceutical research in the formulation development area. The developed models are successful with greater accuracy in predicting the dissolution curves of SR tablets (Hana et al. 2019). DL is one of the most extensively utilized ML methods. DL and ML have revolutionized the world's perspective. Optimizing the formulation of an SR dosage form is a crucial step in pharmaceutical development. To achieve this efficiently, researchers often employ statistical experimental designs, which are valuable techniques for systematically exploring the relationships between formulation variables and the dissolution rate. These help in achieving the desired dissolution times with minimal experimentation.

In the present study, as mentioned in the methodology segment, the developed models have been analyzed by ML techniques to forecast which produces the greatest results. Regarding the outcomes displayed in Table 1, the outputs are as per the 5-fold cross-validation scheme. The ML technique, especially RFR, has shown the best in TPOT AUTOML analysis, but upon hypertuning, the DL model results in greater R^2 and NRMSE%. After the TPOT AUTOML output analysis, we went ahead with the final model development of a 10-fold cross-validation scheme where the DL model was trained over 2200 epochs, fetching us better results than previous ones. Moreover, in Fig. 2, critical parameters that affect the dissolution time were plotted. A clear analysis was conducted to figure out the fundamentals underlying the final model's forecasting by employing Shapley values. The findings not only provide insights into dissolution time prediction but also reinforce the suitability of autoML-based approaches for addressing the challenges posed by complex pharmaceutical tasks. However, the results showed that DL has achieved significant outcomes compared to the other ML models. Therefore, our predictions were consistent (Hana et al. 2018; Yanga et al. 2019).

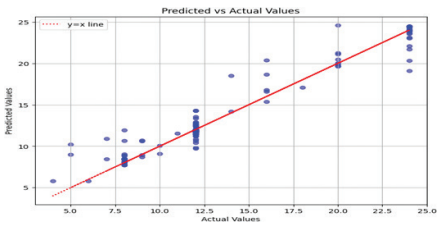

Figure 2. Scatter plot between actual and predictive values for dissolution time of the DL model.

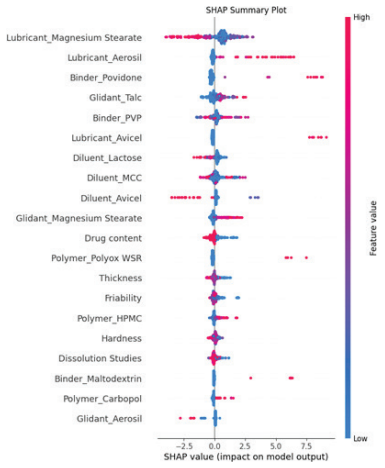

Figure 3. SHAP dependence diagram for the ML models for the top 20 attributes.

Conclusion

SR tablets represent a notable advancement in formulation development, especially when compared to the traditional "trial-and-error" methods that have been in use for

hundreds of years. The traditional approach often involves repeated experimentation and refinement, consuming significant time, financial resources, and human efforts. The present contemporary research has progressed the ML and DL models to exhibit a beneficial estimation for the dissolution time of SR tablet formulations. In summary, the employed models could constructively anticipate the key attributes, but the results showed that predictions with DL were better than those of other ML models. The proposed DL model not only streamlines SR tablet development but also showcases potential applications across different dosage forms and pharmaceutical research areas. Its contribution to reducing development timelines, optimizing resources, and facilitating robust drug product development underscores its significance in advancing pharmaceutical formulation research.

Acknowledgements

The authors express their gratitude to the SRM College of Pharmacy for their insightful recommendations.

References

Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A (2016) Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Molecular Pharmaceutics 13: 2524–2530. https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.6b00248

Bhagat VC, Awate P B, Kardile D P, Pawar O P, Shete RV, Bichukale A D, Raut J M, Mane S R (2023) A review on sustained release drug delivery and sustained release matrix tablets. Journal of Coastal Life Medicine 2: 1174–1190.

Fadel S, Statistics Canada (2022) Explainable Machine Learning Game theory and SHAPLEY Values: A technical review. https://www.statcan.gc.ca/en/data-science/network/explainable-learning

Fink E, Celikovic S, Rehrl J, Sacher S, Urich JAA, Khinast J (2023) Prediction of dissolution performance of uncoated solid oral dosage forms via optical coherence tomography. European Journal of Pharmaceutics and Biopharmaceutics 189: 281–290. https://doi.org/10.1016/j.ejpb.2023.07.003

Galatta DL, Konyves Z, Nagy B, Novak M, Meszaros LA, Szabo E, Farkas A, Marosi G, Nagy ZK (2021) Real-time release testing of dissolution based on surrogate models developed by machine learning algorithms using NIR spectra, compression force, and particle size distribution as input data. International Journal of Pharmaceutics 597:120338. https://doi.org/10.1016/j.ijpharm.2021.120338

Gronberg MP, Beadle BM, Garden AS, Skinner H (2023) Deep learning–based dose prediction for automated, individualized quality assurance of head and neck radiation therapy plans. Practical Radiation Oncology 13: 282–291. https://doi.org/10.1016/j.prro.2022.12.003

Hameed MM, AlOmar MK, Khaleel F, Ansari NA (2021) An extra tree regression model for discharge coefficient prediction: Novel, practical applications in the hydraulic sector and future research directions. Mathematical Problems in Engineering 10: 1–19. https://doi.org/10.1155/2021/7001710

Hana R, Xiong H, Ye Z, Yang Y, Huang T, Jing Q, Lu J, Pan H, Ren F, Ouyang D (2019) Predicting physical stability of solid dispersions by machine learning techniques. Journal of Controlled Release 311: 16–25. https://doi.org/10.1016/j.jconrel.2019.08.030

Hana R, Yang Y, Li X, Ouyang D (2018) Predicting oral disintegrating tablet formulations by neural network techniques. Asian Journal of Pharmaceutical Sciences 13: 336–342. https://doi.org/10.1016/j.ajps.2018.01.003

Hayashi Y, Nakano Y, Marumo Y, Kumada S, Okada K, Onuki Y (2023) A data-driven approach to predicting tablet properties after accelerated test using raw material property database and machine learning. Chemical and Pharmaceutical Bulletin 71: 406–415. https://doi.org/10.1248/cpb.c22-00538

Jiang J, Ma X, Ouyang D, Williams RO (2022) Emerging artificial intelligence (ai) technologies used in the development of solid dosage forms. Pharmaceutics 14: 1–26. https://doi.org/10.3390/pharmaceutics14112257

Loua H, Lian B, Hageman MJ (2021) Applications of machine learning in solid oral dosage form development. Journal of Pharmaceutical Sciences 110: 3150–3165. https://doi.org/10.1016/j.xphs.2021.04.013

Mazur H, Erbrich L, Quodbach J (2023) Investigations into the use of machine learning to predict drug dosage form design to obtain desired release profiles for 3D printed oral medicines. Pharmaceutical development and Technology 28: 219–231. https://doi.org/10.1080/10837450.2023.2173778

Mishra S (2019) Sustained release oral drug delivery system: a concise review. International journal of pharmaceutical sciences Review and research 54: 5–15.

Mohsen A, Tripathi LP, Mizuguchi K (2021) Deep Learning Prediction of Adverse Drug Reactions in Drug Discovery Using Open TG–GATEs and FAERS Databases. Frontiers in drug discovery 1: 1–9. https://doi.org/10.3389/fddsv.2021.768792

Momeni M, Rakhshani S, Abbaspour M, Alizadeh F, Sheikhi N, Zadeh FG, Habibi Z, Tabesh H (2023) Dataset development of pre-formulation tests on fast disintegrating tablets (FDT): data aggregation. BMC Research Notes 16: 1–5. https://doi.org/10.1186/s13104-023-06416-w

Santoshrao PH, Dhilipbhai PN, Vinay PS, Rane BR, Gujarathi NA, Pawar SP (2014) A Review on Sustained Release Drug Delivery System. International Journal of Pharmaceutical, Chemical and Biological Sciences 4: 470–478.

Sustersic T, Gribova V, Nikolic M, Lavalle P, Filipovic N, Vrana NA (2023) The Effect of Machine Learning Algorithms on the Prediction of Layer-by-Layer Coating Properties. ACS Omega 8: 4677–4686. https://doi.org/10.1021/acsomega.2c06471

Szlek J, Khalid MH, Pacławski A, Czub N, Mendyk A (2022) Puzzle out Machine Learning Model-Explaining Disintegration Process in ODTs. Pharmaceutics 14: 1–23. https://doi.org/10.3390/pharmaceutics14040859

Takayama K, Kawai S, Obata Y, Todo H, Sugibayashi K (2017) Prediction of Dissolution Data Integrated in Tablet Database Using Four-Layered Artificial Neural Networks. Chemical and Pharmaceutical Bulletin 65: 967–972. https://doi.org/10.1248/cpb.c17-00539

Wang H, Kwong CF, Liu K, Liu Z, Chen Z (2022) A Novel Artificial Intelligence System in Formulation Dissolution Prediction. Computational intelligence and neuroscience 2022: 1–11. https://doi.org/10.1155/2022/8640115

Yanga Y, Yea Z, Sua Y, Zhao Q, Lib X, Ouyang D (2019) Deep learning for *in vitro* prediction of pharmaceutical formulations. Acta Pharmaceutica Sinica B 9: 177–185. https://doi.org/10.1016/j.apsb.2018.09.010