Article

Not peer-reviewed version

Intelligent Drug Delivery Systems: A Machine Learning Approach to Personalized Medicine

Yi Wang *, Wenxuan Shao , JungHua Lin , Shirong Zheng

Posted Date: 6 May 2025

doi: 10.20944/preprints202504.2570.v1

Keywords: personalized drug delivery system; machine learning; precision medicine; patient data analysis

 doi:10.20944/preprints202504.2570.v1

Article

Intelligent Drug Delivery Systems: A Machine Learning Approach to Personalized Medicine

Yi Wang [1,*], Wenxuan Shao [1], Junghua Lin [2] and Shirong Zheng [3]

[1] Hebei University of Economics and Business, Hebei, China
[2] Suffolk university, USA
[3] Purdue University, West Lafayette, USA
* Correspondence: 2922417923@qq.com

Abstract: This study proposes a novel framework for personalized drug delivery by leveraging machine learning techniques. Using a dataset of 10,000 patient records, we developed and evaluated three ensemble models—XGBoost, LightGBM, and CatBoost—to predict optimal drug delivery parameters based on individual characteristics.The dataset includes diverse attributes such as demographics, medical conditions, treatment history, and clinical outcomes, providing a solid foundation for personalized medicine. We performed extensive data preprocessing and feature engineering, followed by the implementation and comparison of the three machine learning algorithms. Results indicated that XGBoost achieved the best overall performance (accuracy = 0.6386, F1 = 0.6275), while LightGBM attained the highest recall (0.6578).Model performance was assessed using multiple metrics—accuracy, precision, recall, and F1 score—with particular attention to convergence and learning curves. These findings suggest that machine learning can effectively capture complex patterns in patient data to support personalized drug delivery.While the current models yield promising results, they highlight opportunities for improvement through larger datasets and more advanced algorithms. This work contributes to the evolving field of precision medicine by offering a quantitative framework to optimize drug delivery based on individual characteristics.

Keywords: personalized drug delivery system; machine learning; precision medicine; patient data analysis

1. Introduction

The advent of precision medicine has transformed healthcare delivery, with personalized drug delivery systems (PDDS)assuming a critical role in modern medical practice. Conventional drug delivery systems often adopt a "one-size-fits-all" approach, disregarding individual patient variability, which may lead to suboptimal therapeutic outcomes or unnecessary side effects. The rapid advancement of machine learning techniques has introduced novel strategies to address these challenges.

This study aims to develop a machine learning-based prediction model for personalized drug delivery. By analyzing multidimensional medical data-including patient demographics, clinical indicators, and treatment records, combined with modern machine learning algorithms, we seek to customize optimal drug delivery strategies for individual patients. This approach not only enhances therapeutic efficacy but also reduces the risk of adverse reactions, thereby enhancing the overall quality of healthcare.

The primary objectives of this study are to: (1) establish a reliable patient data analysis framework, (2) develop and optimize machine learning models for predicting personalized drug release curves, and (3) evaluate the performance of different machine learning algorithms in this application.

 doi:10.20944/preprints202504.2570.v1

2. Literature Review

The integration of artificial intelligence (AI) and machine learning into drug delivery systems represent a significant advancement in pharmaceutical technology. The literature review reveals several key developments and perspectives in this field.

Kamaly et al. provided a comprehensive review of degradable controlled-release polymers and polymeric nanoparticles, establishing the fundamental mechanisms of drug release control [1]. Their work highlighted the importance of understanding polymer-drug interactions and release kinetics, which serves as a foundation for developing intelligent drug delivery systems. They particularly emphasized how different polymer architectures and properties can be manipulated to achieve desired drug release profiles, which is crucial for personalized medicine applications.

Building on these fundamentals, Schneider et al. presented a paradigm shift in drug design through the lens of artificial intelligence [2]. Their research demonstrated how AI technologies could revolutionize traditional drug development approaches, particularly in predicting drug-target interactions and optimizing delivery parameters. They proposed that machine learning algorithms could significantly reduce the time and cost associated with drug development while improving the accuracy of delivery predictions.

Hassanzadeh et al. further explored the significance of artificial intelligence in drug delivery system design [3]. Their work emphasized how AI could enhance the development of smart drug delivery systems by optimizing various parameters such as particle size, drug loading efficiency, and release kinetics. They particularly highlighted the potential of machine learning algorithms in predicting drug release patterns and personalizing treatment regimens based on patient-specific factors.

Recent developments, as documented by Gholap et al., have shown remarkable progress in applying AI to drug delivery and development [4]. Their comprehensive review detailed how various machine learning techniques, including deep learning and ensemble methods, can be effectively utilized to design more efficient drug delivery systems. They specifically addressed how AI can help in predicting drug-polymer compatibility and optimizing formulation parameters.

Serrano et al. focused on the revolutionary impact of AI applications in personalizing medicine through improved drug discovery and delivery methods [5]. Their research demonstrated how machine learning algorithms could analyze patient-specific data to optimize drug delivery parameters, potentially leading to more effective and safer treatments. They emphasized the role of AI in developing predictive models that can account for individual patient variations in drug response.

Vora et al. provided insights into the practical applications of AI in pharmaceutical technology and drug delivery design [6]. Their work highlighted how machine learning algorithms could be used to predict drug release profiles and optimize delivery system parameters. They particularly emphasized the potential of AI in developing smart drug delivery systems that can adapt to patient-specific needs and conditions.

Collectively, these studies illustrate the evolving landscape of AI applications in drug delivery systems, highlighting both the progress made and the potential for future developments. The literature indicates that machine learning approaches can significantly improve the efficiency and effectiveness of drug delivery systems while moving towards more personalized treatment approaches. However, it also indicates that continued research is needed to fully realize the potential of AI in this field, particularly in validating predictive models and implementing them in clinical settings.

This comprehensive body of research provides a strong foundation for our current study, which aims to develop and implement machine learning models for predicting personalized drug delivery profiles. The existing literature supports our approach while highlighting areas where our research can contribute to advancing the field further.

 doi:10.20944/preprints202504.2570.v1

3. Data Introduction

This study uses patient data from medical data sets for analysis. The data set contains 10,000 patient records, covering several key medical characteristic variables. Specifically, the data set contains important indicators such as basic demographic characteristics of patients (such as age and gender), medical insurance type, diagnosis results, hospitalization information (including hospitalization type and length of stay) and medical expenses. All patient data are anonymized to protect patient privacy. The disease diagnosis in the data set covers many common diseases, including cardiovascular diseases, respiratory diseases, digestive system diseases, etc., and has strong representativeness and universality.

Before analyzing the original dataset, we first performed data cleaning, including outlier removal, missing value processing, and other steps, to remove erroneous data and improve the accuracy and reliability of subsequent analysis.

In order to ensure the data quality, this paper preprocesses the original data set, including missing value processing, abnormal value detection and processing. In the process of data cleaning, this paper pays special attention to the distribution characteristics of numerical variables (such as medical expenses and hospital stay), and codes and standardizes classified variables (such as insurance types and diagnosis results). After data preprocessing, a complete and reliable analysis sample is finally obtained, which provides a solid data foundation for the training and verification of the subsequent machine learning model.

The dataset's key advantage lies in its rich clinical and treatment-related information, making it highly valuable for developing personalized drug delivery models.By analyzing this multidimensional patient data, we can explore the relationship between individual characteristics and treatment outcomes, thereby supporting the optimization of drug delivery systems.

Table 1 shows the variable explanation of the data set.

Table 1. Variables and descriptions.

Variable Name	Variable Type	Variable Description
Patient Name	Text	Name of the patient
Age	Numerical	Age of the patient at admission (in years)
Gender	Categorical	Gender of the patient (Male/Female)
Blood Type	Categorical	Blood type of the patient (e.g., A+, O-)
Medical Condition	Categorical	Primary diagnosis of the patient (e.g., Diabetes, Hypertension, Asthma)
Date of Admission	Date	Date when the patient was admitted to the healthcare facility
Doctor	Text	Name of the doctor responsible for the patient's care
Hospital	Text	Name of the healthcare facility where the patient was admitted
Insurance Provider	Categorical	Insurance provider of the patient (e.g., Aetna, Blue Cross)
Billing Amount	Continuous	Cost of healthcare services during the patient's admission (in USD)
Room Number	Text	Room number where the patient stayed during admission

 doi:10.20944/preprints202504.2570.v1

Admission Type	Categorical	Type of admission (Emergency/Elective/Urgent)
Discharge Date	Date	Date when the patient was discharged from the healthcare facility
Medication	Categorical	Medications prescribed or administered to the patient (e.g., Aspirin, Ibuprofen)
Test Results	Categorical	Results of medical tests conducted (Normal/Abnormal/Inconclusive)

Figure 1 shows the number distribution of male and female patients with different blood types. Through this chart, we can intuitively observe the difference in the number of patients of different sexes in each blood group category. From the overall distribution, the number of male and female patients in each blood type category fluctuates to some extent, which indicates that there may be some potential correlation between blood type and gender. For example, in some blood types (e.g., A+, B+, etc.), a noticeable disparity exists between male and female patient counts. This may suggest that gender plays a role in blood group distribution, potentially due to underlying genetic or physiological factors.

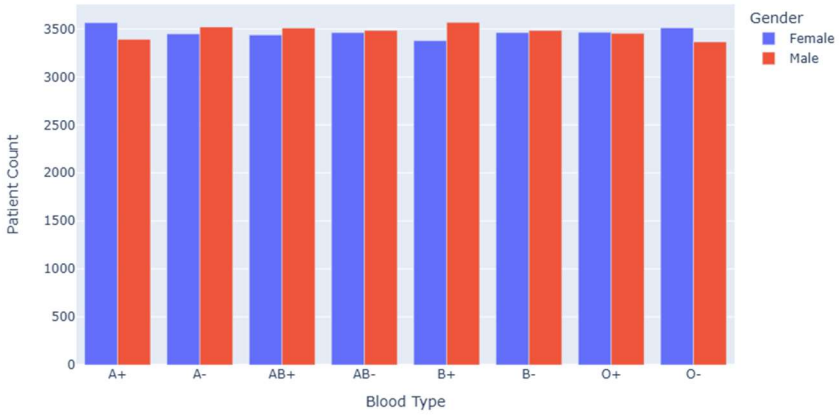

Figure 1. Distribution of Patient Counts by Blood Type and Gender.

Figure 2 shows the number distribution of patients with various diseases under different blood types. The figure shows the number of patients with arthritis, asthma, cancer, diabetes, hypertension, obesity and other diseases in different blood types. As can be seen from the figure, the number of patients with different diseases in each blood type is uneven. Taking the A+blood group as an example, the number of patients suffering from diseases such as diabetes and hypertension is relatively large; In the B+blood group, the number of patients with asthma and obesity may be more prominent. This shows that there may be some relationship between blood type and disease type, and people with certain blood types may have higher susceptibility to specific diseases.

 doi:10.20944/preprints202504.2570.v1

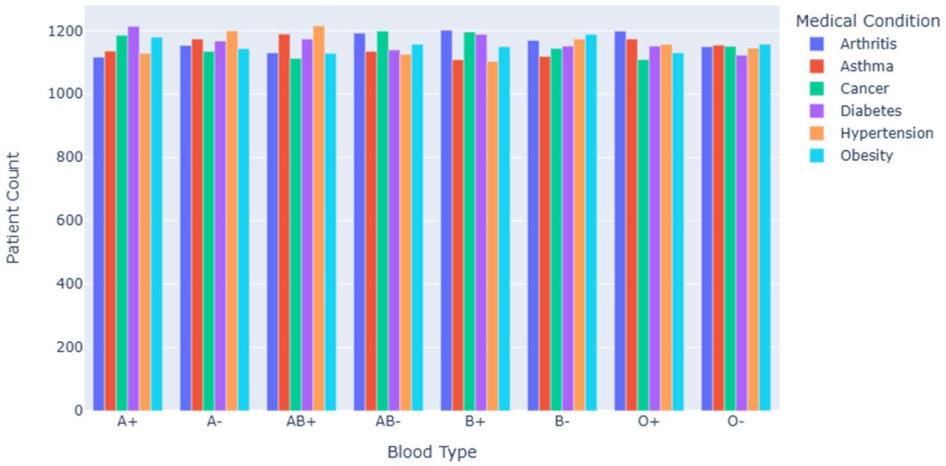

Figure 2. Patient Count by Blood Type and Medical Condition.

Figure 3 illustrates the distribution of patients with various medical conditions across different admission types, including emergency and elective admissions. The data reveal distinct differences in admission patterns among disease categories.

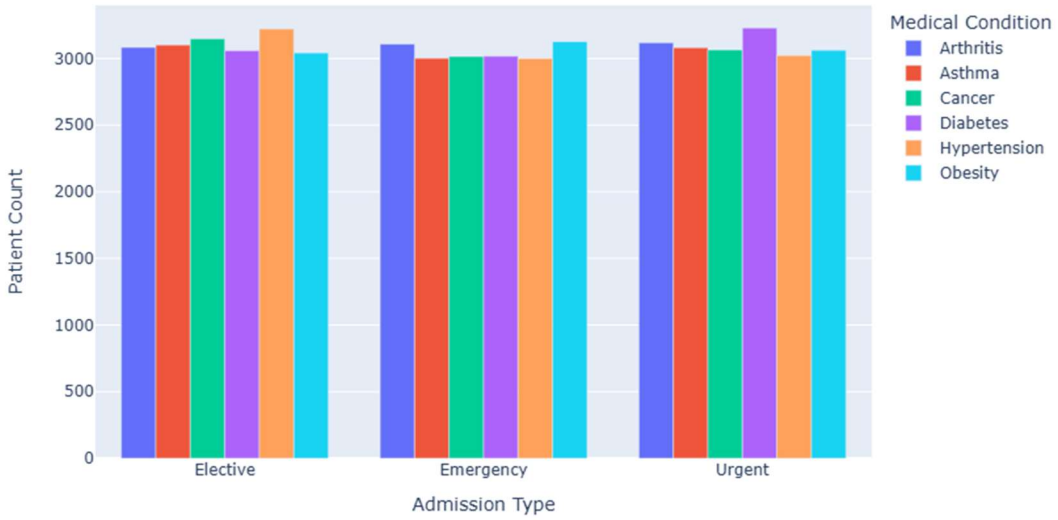

Figure 3. Distribution of Patient Counts by Admission Type and Medical Condition.

For instance, patients with arthritis are predominantly admitted on an elective basis, while patients with asthma and cancer are more frequently admitted as emergencies. This trend reflects the clinical urgency associated with different conditions.

Diseases such as asthma and cancer often have acute onset and require immediate medical intervention, resulting in a higher rate of emergency admissions. In contrast, arthritis is generally more stable, allowing patients to schedule hospitalizations based on their treatment plans and personal convenience.

 doi:10.20944/preprints202504.2570.v1

4. Model Introduction

4.1. XGBoost (eXtreme Gradient Boosting)

XGBoost is an efficient, distributed gradient boosting decision tree algorithm that improves upon traditional GBDT in both algorithmic design and engineering implementation [7,8]. Its core objective function comprises two components: a training loss term and a regularization term.

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

where $l(y_i, \hat{y_i})$ is the training loss function, which measures the difference between the predicted value and the real value; $\Omega(f_k)$ is a regularization term used to control the complexity of the model. The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda |w|^2$$

Here, γT controls the number of leaf nodes, and λ controls L2 regularization of leaf weights. In the t-round iteration, the objective function can be approximated by the second-order Taylor expansion as:

$$L^{(t)} = \sum_{i=1}^{n} \left[g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where g_i and h_i are the first and second derivatives of the loss function respectively.

4.2. LightGBM (Light Gradient Boosting Machine)

LightGBM is an efficient gradient lifting framework developed by Microsoft. Its main innovation lies in the decision tree algorithm based on histogram and the leaf growth strategy with depth restriction [9]. When the node is split, the gain calculation formula is:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

where G_L and G_R are the first-order statistics of the left and right child nodes, and H_L and H_R are the second-order statistics. The formula for calculating the optimal leaf value is:

$$w^* = -\frac{G}{H + \lambda}$$

LightGBM greatly reduces the memory consumption and calculation amount through histogram algorithm, and adopts unique feature parallelism and data parallelism methods to improve the training speed.

4.3. CatBoost (Categorical Boosting)

CatBoost is a machine learning algorithm developed by Yandex, which especially optimizes the processing of category features and introduces sorting promotion to prevent prediction deviation [10,11]. Its prediction function is:

$$\hat{y_i} = \sum_{j=1}^{n} a_j \cdot \text{Tree}_j(x_i)$$

 doi:10.20944/preprints202504.2570.v1

For category features, CatBoost uses an innovative coding method of target statistics:

$$\text{CategoryAvg} = \frac{\sum_{i=1}^{n}[x_i = \text{category}] \cdot y_i + a \cdot P}{\sum_{i=1}^{n}[x_i = \text{category}] + a}$$

WHERE a is the smoothing parameter and p is the prior probability. The overall loss function of the model is:

$$L(y, f) = \sum_{i=1}^{n} l\big(y_i, f(x_i)\big) + \sum_{j=1}^{n} \lambda |f_j|^2$$

These three ensemble learning algorithms each exhibit unique characteristics and offer distinct advantages in practical applications.

The primary strengths of the XGBoost algorithm lie in three key areas. First, it improves model convergence by approximating the objective function through second-order Taylor expansion. Second, it incorporates a regularization term to control model complexity and reduce the risk of overfitting. Third, XGBoost supports both feature-parallel and data-parallel computation, significantly enhancing computational efficiency.

As a lightweight gradient boosting framework, LightGBM is particularly advantageous in terms of computational performance. It introduces a histogram-based decision tree algorithm that reduces memory usage. Additionally, the leaf-wise growth strategy minimizes computation, and its efficient parallel processing further accelerates training speed.

CatBoost, on the other hand, demonstrates unique advantages in handling categorical features. It introduces an innovative approach for processing categorical variables, improving model performance on such data. Furthermore, it reduces prediction bias through the implementation of ordered boosting. CatBoost also automatically handles missing values and categorical features, thus simplifying data preprocessing and reducing the manual workload.

5. Model Results Analysis

The test results of three models on the dataset after data cleaning are shown in Table 2. According to the data in Table 2, the XGBoost classifier achieved an accuracy of 0.6386, precision of 0.6175, recall of 0.6456, and an F1 score of 0.6275. The LGBM classifier yielded an accuracy of 0.6015, precision of 0.6023, recall of 0.6578, and an F1 score of 0.6235. The CatBoost classifier reported an accuracy of 0.3355, precision of 0.5946, recall of 0.6386, and an F1 score of 0.6175.

Table 2. Comparison of classification results of different models.

Model	Accuracy	Precision	Recall	F1
XGBoost Classifier	0.6386	0.6175	0.6456	0.6275
LGBM Classifier	0.6015	0.6023	0.6578	0.6235
CatBoost Classifier	0.3355	0.5946	0.6386	0.6175

Overall, the performance metrics of the three models are relatively close. XGBoost demonstrated slightly higher accuracy and F1 score than the other two models, indicating better overall performance. LGBM achieved the highest recall, suggesting a stronger ability to capture positive instances. In contrast, CatBoost performed slightly weaker across all metrics.

These indicators collectively reflect the models' classification performance. Accuracy measures the proportion of all correct predictions, precision represents the proportion of correctly predicted positive samples, recall reflects the proportion of actual positives correctly identified, and the F1 score provides a balanced evaluation of both precision and recall.

The training loss curve in Figure 4 depicts the variation of loss value with the number of iterations. In the initial stage, the loss value is about 1.1 when the number of iterations is close to zero, and the model fitting effect is not good. With the progress of iteration, the loss value in the first 200

 doi:10.20944/preprints202504.2570.v1

times decreased significantly, the model learned quickly, and the performance improved obviously. After that, although the decline rate slowed down, the loss value continued to decrease to about 0.58 by 1000 iterations. The overall curve is monotonically decreasing, which shows that the model can effectively learn and reduce errors in the training process, and it is in an ideal state of convergence, but attention should be paid to its performance in the verification set to prevent over-fitting.

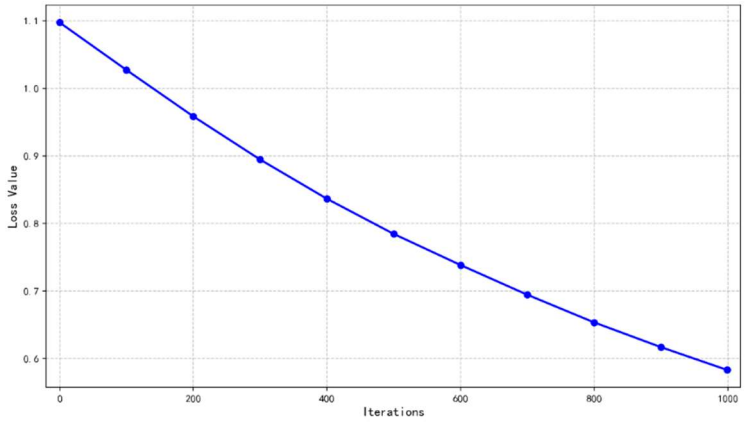

Figure 4. Model Training loss curve.

6. Conclusions

This research contributes to the field of personalized drug delivery by developing and implementing machine learning-based predictive models.A comprehensive analysis of patient data, combined with advanced machine learning algorithms, demonstrates substantial potential for improving drug delivery optimization.The comparison of XGBoost, LightGBM, and CatBoost revealed varied strengths in prediction accuracy and recall, with XGBoost exhibiting marginally superior overall performance in accuracy and F1 score.

The training process showed robust convergence, with the loss function consistently decreasing across iterations and stabilizing at an acceptable level.

However, the current accuracy rates suggest room for improvement, potentially through the incorporation of larger datasets and more sophisticated algorithmic approaches. The findings underscore the viability of machine learning applications in personalized medicine while highlighting areas for future enhancement.

Looking ahead, this research opens promising avenues for future exploration, such as expanding data sources, applying deep learning architectures, and conducting clinical validation.

The implications of this work extend beyond theoretical frameworks, offering practical insights for the advancement of precision medicine and personalized therapeutic approaches. Although challenges remain in achieving optimal prediction accuracy, this study provides a solid foundation for ongoing development in personalized drug delivery.

References

1. Kamaly N, Yameen B, Wu J, et al. Degradable controlled-release polymers and polymeric nanoparticles: mechanisms of controlling drug release[J]. Chemical reviews, 2016, 116(4): 2602-2663.
2. Schneider P, Walters W P, Plowright A T, et al. Rethinking drug design in the artificial intelligence era[J]. Nature reviews drug discovery, 2020, 19(5): 353-364.
3. Hassanzadeh P, Atyabi F, Dinarvand R. The significance of artificial intelligence in drug delivery system design[J]. Advanced drug delivery reviews, 2019, 151: 169-190.
4. Gholap A D, Uddin M J, Faiyazuddin M, et al. Advances in artificial intelligence in drug delivery and development: A comprehensive review[J]. Computers in Biology and Medicine, 2024: 108702.
5. Serrano D R, Luciano F C, Anaya B J, et al. Artificial intelligence (AI) applications in drug discovery and drug delivery: Revolutionizing personalized medicine[J]. Pharmaceutics, 2024, 16(10): 1328.

 doi:10.20944/preprints202504.2570.v1

6. Vora L K, Gholap A D, Jetha K, et al. Artificial intelligence in pharmaceutical technology and drug delivery design[J]. Pharmaceutics, 2023, 15(7): 1916.
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
8. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015, 1(4): 1-4.
9. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.
10. Dorogush A V, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support[J]. arXiv preprint arXiv:1810.11363, 2018.
11. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.