Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Personalized anti-tumor drug efficacy prediction based on clinical data

Xinping Xie [a], Dandan Li [a], Yangyang Pei [a], Weiwei Zhu [b], Xiaodong Du [c], Xiaodong Jiang [d], Lei Zhang [e], Hong-Qiang Wang [b],*

[a] *School of Mathematics and Physics, Anhui Jianzhu University, Hefei, China*
[b] *Institute of Intelligent Machines/Zhongqi AI Lab., Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China*
[c] *Experimental Teaching Center, Hefei University, Hefei, China*
[d] *Medical Oncology Department, The First Affiliated Hospital of University of Science and Technology of China, Hefei, Anhui, 230001, China*
[e] *Pharmacy Department, The First Affiliated Hospital of University of Science and Technology of China, Hefei, Anhui, 230001, China*

ARTICLE INFO

Keywords:
Drug efficacy prediction
Clinical data
Text mining
Clinical decision-making

ABSTRACT

Anti-tumor drug efficacy prediction poses an unprecedented challenge to realizing personalized medicine. This paper proposes to predict personalized anti-tumor drug efficacy based on clinical data. Specifically, we encode the clinical text as numeric vectors featured with hidden topics for patients using Latent Dirichlet Allocation model. Then, to classify patients into two classes, responsive or non-responsive to a drug, drug efficacy predictors are established by machine learning based on the Latent Dirichlet Allocation topic representation. To evaluate the proposed method, we collected and collated clinical records of lung and bowel cancer patients treated with platinum. Experimental results on the data sets show the efficacy and effectiveness of the proposed method, suggesting the potential value of clinical data in cancer precision medicine. We hope that it will promote the research of drug efficacy prediction based on clinical data.

1. Introduction

Cancer, as malignant tumor, has become one of the leading causes of death in the global population [1]. The numbers of new cases and deaths of cancer every year remain high over past decades [2]. Conventional cancer treatment strategies, which typically treat the same disease with the same treatment, often result in unsatisfactory outcomes with severe toxic side effects in patients due to individual variability [3]. Therefore, it has become a clinically urgent need to predict drug efficacy in individual cancer patients for precision treatment.

Currently, most of drug efficacy prediction researches are based on expensive genomics data. As we know, two international large-scale research projects, Cancer Cell Line Encyclopedia (CCLE) [4] and Genomics of Drug Sensitivity in Cancer (GDSC) [5], have been realized and released genomics data, such as transcription and methylation profiles, with sensitivity assays to drugs. Following this, Li et al. [6] used gene expression data to do “sensitive-resistant” binary prediction of cell lines to anti-tumor drugs with the help of Support Vector Machines (SVMs). Bai et al. [7] studied how to combine gene sequencing data and drug information and developed a dual input and dual output deep convolutional neural network (CNN) to predict drug sensitivity. Li et al. [8] proposed DeepDSC model,

* Corresponding author.
E-mail addresses: 592745039@qq.com (X. Xie), 2117605767@qq.com (D. Li), 466691678@qq.com (Y. Pei), juway@126.com (W. Zhu), xdongdu@hfuu.edu.cn (X. Du), 63561783@qq.com (X. Jiang), zhanglei6@126.com (L. Zhang), hqwang126@126.com (H.-Q. Wang).

https://doi.org/10.1016/j.heliyon.2024.e27300
Received 6 March 2023; Received in revised form 27 February 2024; Accepted 27 February 2024
Available online 4 March 2024

which takes gene expression and chemical features of drugs together as inputs to regress the half maximal inhibitory concentration (IC50) value in cells. In addition, Fang et al. [9] designed a quantile regression forest-based method for drug efficacy prediction and ran it on multi-omics data including gene expression, mutation status, and copy number variations. Although these works have made progress, nowadays, drug efficacy prediction still remains unsolved in clinic due to complex clinical and/or medical factors.

Compared with genomic data, clinic data, including patients' biological attributes, clinical and pathological features, chemical tests, radiology examination, *etc.*, are inexpensive and very large in amount. Especially, as imaging technology, including Computed tomography (CT), B-mode ultrasound (BU), Magnetic Resonance Imaging (MRI), *etc.,* has advanced rapidly in both imaging capability and data quality, radiology examination is playing an increasingly important role in tumor diagnosis and treatment [12]. Generally, they can provide rich morphological and pathological information of tumors in patients, which greatly help understand the patient's lesion condition and treatment response pattern. For example, CT imaging, widely used during the detection and diagnosis of tumors, visualizes the lesion and surrounding structures of cancer tissues, and the range of applications covers various organ systems throughout the body. Wang et al. [13] analyzed CT images to predict the prognosis of EGFR-TKI treatment and screened out the benefit population of targeted therapy. MRI is equally widely used like CT, and its application covers tumors mainly in various tissue systems, such as head, thorax, abdomen and pelvis. Compared with CT, MRI is especially useful in quantitatively understanding the functional and compositional changes of lesions. Zhu et al. [14] extracted high-dimensional features from MRI images and, with the help of machine learning, probed into the overall heterogeneity of tumors. Results demonstrated the usage of MRI image in predicting the efficacy and prognosis of neoadjuvant chemotherapy in cancer. BU is featured with probing the superficial and abdominopelvic region, and by scanning with many angles, it can identify cystic solid lesions with high diagnostic sensitivity [15]. Recently, Dercle et al. [11] analyzed radiomic signatures by machine learning for predicting the response and survival of metastatic colorectal cancer patients treated with agents targeting the EGFR pathway, achieving satisfactory prediction performance.

Clinically, the vast majority of the imaging-based examination results are recorded in electronic medical records (EMRs) in an unstructured text format. As we known, EMRs have become widely available in hospitals these years, which have led to the generation of clinical data (including the imaging examination reports) over the past decades [10]. Generally, the data are directly delivered to physicians for manually making clinical diagnosis and treatment decision on cancer patients. Especially, the radiomic examination reports, including CT, MRI and BU ones, summarized by medical technicians after reading and understanding the radiomic images, are enriched with the medical knowledge and experience from doctors. Logically, such high-level semantic medical information would benefit cancer precision medicine. We argue that the radiomic examination reports could provide an alternative low-cost way to predict personalized anti-tumor drug efficacy.

In this paper, we propose to predict drug efficacy based on clinical text data and developed a machine learning method for anti-tumor efficacy prediction. Specifically, the method mainly relies on mining the potential drug efficacy-associated patterns underlying patients' clinical text data of radiological examination. Generally, there are many text mining methods, e.g., association rule analysis [16], clustering [17], *etc.*, most of which, however, are deficient in processing irregular clinical texts [18]. Recently, powerful topic models, e.g. Latent Dirichlet Allocation (LDA) [19], are developed featured with mining abstract topics underlying text data. Compared with other methods, it can efficiently explore relationships hidden in complex text data by overcoming the problems of semantic loss and data sparsity. In recent years, topic models have been used to uncover patient treatment patterns in clinical text for some diseases. For example, Rumshisky et al. [20] trained an LDA model with 75 topics to predict the probability of readmission for early psychiatric patients, and Zalewski et al. [21] applied topic model to assess health status for inpatients. Li et al. [22] developed supervised topic models to predict antidepressant treatment stability. Similarly, we here employ the LDA model to analyze clinical text data for extracting personalized disease features associated with treatment response and then construct drug efficacy machine learning (ML) predictors based on the resulting low-dimensional representation for personalized cancer medicine.

To evaluate the proposed method, we collected from our collaborated hospital (The First Affiliated Hospital of University of Science and Technology of China) and collated clinical records (2016–2022) of cancer patients treated with platinum. The experimental results on the data set show that the proposed method achieved the competitive performance of drug efficacy prediction, suggesting the

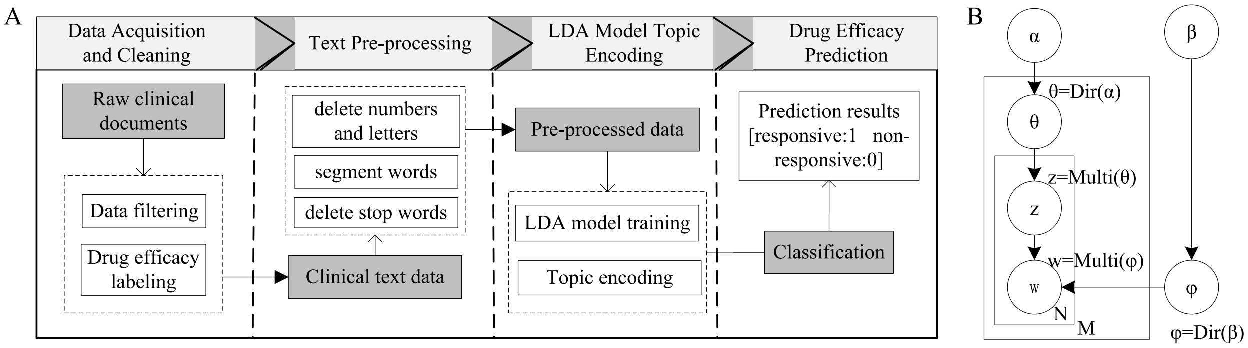

Fig. 1. Overall workflow of the proposed method (A) and the Bayesian network diagram of LDA model (B). α, the hyperparameter of the prior Dirichlet distribution for each topic distribution; β, the hyperparameter of the prior Dirichlet distribution for each topic word distribution; θ, the probability distribution of topics; φ, the probability distribution of words; z, the hidden topics; w, the observable words in the document.

potential value of clinical application. The rest of the paper is organized as follows: Section 2 presents the details of the proposed method, including the encoding of patient clinical text data based on LDA model, the drug efficacy prediction model and the evaluation metrics used. The experimental results are given in Section 3. Finally, the paper is concluded in Section 4.

2. Methods

Fig. 1A shows the flowchart of the drug efficacy prediction method based on clinical text data. We cleaned up the raw clinical data of cancer patients collected from the hospitals, including data filtering and patient screening, and labeled patients with being responsive or non-responsive to cisplatin. Specifically, we only kept patients with drug efficacy assessment, and screened patients using following criteria: (i) patients with NSCLC and who had received first-line platinum-based chemotherapy; (ii) patients with treatment outcome available after one or two courses of treatment. We labeled patients according to the RECIST criteria: CR, PR and SD are classified as responsive (1) and PD as non-responsive (0). For the qualified patients, the text data were pre-processed [23] by 1) deleting numbers and letters, 2) segmenting words and 3) deleting stop words. After the pre-processing, we built LDA model by extracting latent topics and encoded the text documents for each patient. Finally, with the LDA representations of patients, we employed machine learning models, including Logistic Regression (LR), k-Nearest Neighbor (KNN), Decision Tree (DT) and SVMs, to predict the class of drug efficacy in patients. Essentially, the method mainly relies on how well to uncover the complex hidden radiological text patterns associated with treatment outcomes.

2.1. *Ethics*

The study was performed in adherence with the Declaration of Helsinki and its later amendments. The study was approved by the Ethics Committee of The First Affiliated Hospital of University of Science and Technology of China (approval no. 2021-RE-85), and the requirement for written informed consent was waived, and the personal identifiers were removed before the data analysis.

2.2. *Text mining and LDA model*

LDA model is commonly used in natural language processing and text mining tasks, providing a good way to mine valuable information underlying massive text data. It can not only uncover the latent topics in document corpus by disentangling the potential semantic relationship between documents, but also effectively reduce data dimensionality to alleviate the problem of data sparsity. The topic model has been extensively applied to many fields such as sentiment analysis [24], topic evolution [25], text classification [26] and biomedicine [27].

Essentially, LDA, proposed by David Blei et al. [28], can be seen as a probabilistic topic model for text mining. The basic idea behind it is to take a document as a combination of different latent topics, each of which represents a specific distribution of words [29]. Mathematically, it mainly defines four variables: document-topic distribution θ, topic-word distribution φ, latent topic vector z and observable word vector w, and the probability of the observed w occurring in a document d can be written as $p(w|d) = p(z|d)p(w|z)$. So the LDA model can be seen as a Bayesian network with a three-layer structure: documents, topics and words, as shown in Fig. 1B.

2.3. *Radiological text data encoding based on LDA*

In order to characterize the lesion and pathological condition of patients before treatment, we take the radiology examination text as a document for analysis. Assuming M patients, N observable words and K potential topics, we denote the set of patient documents as $D = \{W_1, W_2, ..., W_M\}$, where $W_i = \{w_1, w_2, ..., w_N\}, i = 1, 2, \cdots, M$, denotes the word vector of the ith patient document, and the topic set as $Z = \{z_1, z_2, ..., z_K\}$, where z_k $(k = 1, 2, \cdots, K)$ denotes the kth topic.

According to the dependence shown in Fig. 1B, the joint probability distribution for a patient document W can be formulated as

$$p(\theta, Z, \varphi, W|\alpha, \beta) = p(\theta|\alpha)\prod_{k=1}^{K} p(\varphi_k|\beta)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \varphi_{z_n}) \tag{1}$$

Then, from Eq. (1), by means of integration and summation, the generation probability of W can be written as

$$p(W|\alpha, \beta) = \iint p(\theta|\alpha)\prod_{k=1}^{K} p(\varphi_k|\beta)\left(\prod_{n=1}^{N}\sum_{z_{dn}} p(z_n|\theta)p(w_n|z_n, \varphi_{z_n})\right)d\theta d\varphi_k \tag{2}$$

One can solve for the parameters α and β by maximizing the log-likelihood function $logp(W|\alpha, \beta)$ of Eq. (2). Considering the coupling of the latent variables θ, Z and φ, we use variational inference method [30] to approximate the true posterior distributions $p(\theta, Z, \varphi|W, \alpha, \beta)$ by introducing a variational distribution $q(\theta, Z, \varphi|\gamma, \varphi, \lambda)$. The objective function of $logp(W|\alpha, \beta)$ can be further written as

$$\log p(W|\alpha, \beta) = L(\gamma, \varphi, \lambda; \alpha, \beta) + D(q(\theta, Z, \varphi|\gamma, \varphi, \lambda)||p(\theta, Z, \varphi|W, \alpha, \beta)) \tag{3}$$

where $L(\gamma, \varphi, \lambda|\alpha, \beta)$ denotes the lower bound on the log-likelihood and $D(q||p)$ denotes the Kullback-Leibler divergence measuring the similarity of the two distributions.

The EM algorithm [31] is next used to solve for the optimal parameters and latent variables. Following Eq. (3), we first use

$L(\gamma,\varphi,\lambda|\alpha,\beta)$ to approximate the $logp(W|\alpha,\beta)$ and solve for the optimal variational parameters γ,φ,λ by maximizing $L(\gamma,\varphi,\lambda|\alpha,\beta)$ in the E step. The variational parameters γ,φ,λ are updated according to Eqs. (4)–(6).

$$\varphi_{nk}\propto\exp\left\{\psi(\gamma_k)-\psi\left(\sum_{k'=1}^{K}\gamma_{k'}\right)+\sum_{v=1}^{V}w_{nv}\left[\psi(\lambda_{kv})-\psi\left(\sum_{i=1}^{V}\lambda_{ki}\right)\right]\right\} \tag{4}$$

$$\gamma_k=\alpha_k+\sum_{n=1}^{N}\varphi_{nk} \tag{5}$$

$$\lambda_{kv}=\beta_v+\sum_{d=1}^{M}\sum_{n=1}^{N_d}\varphi_{dnk}w_{dnv} \tag{6}$$

Then, the model parameters α and β are updated in the M step. We repeat steps E and M until convergence. Since φ_{nk} denotes the probability that the nth word in the document W is generated by topic k, it is deduced that $\sum_{n=1}^{N}\varphi_{nk}$ is an estimate of the number of words generated by topic k in the document W. Then we can get the topic distribution θ_M of W as Eq. (7).

$$\theta_M=\left[\frac{\sum_{n=1}^{N}\varphi_{n1}}{N},\dots,\frac{\sum_{n=1}^{N}\varphi_{nk}}{N},\dots,\frac{\sum_{n=1}^{N}\varphi_{nK}}{N}\right] \tag{7}$$

In addition, we use perplexity as the criterion to optimize the number of hidden topics and determine the optimal number of topics by plotting the perplexity-topic curve. The perplexity is calculated as

$$perplexity(D)=\exp\left(-\frac{\sum_{i=1}^{M}\log p(w_i)}{\sum_{i=1}^{M}N_i}\right) \tag{8}$$

where N_i denotes the total number of words in the ith text document, and $p(w_i)$ denotes the probability of the word vector w_i of the ith text document. The lower the perplexity of Eq. (8) the better the model learned [32].

2.4. Drug efficacy prediction model

We view the drug efficacy prediction task as a binary classification problem: responsive (1) or non-responsive (0). Based on the topic features extracted from the LDA model, we employ the following binary ML classifiers for extensive evaluation: LR, KNN, DT and SVMs, all of which are popular and competitive machine learning algorithms used commonly in practice [33]. Especially, they are simple, explainable and use-friendly, and thus more suitable for solving clinical medicine problems. Although these general models may not exhibit extra strength in text processing, the application of LDA before them can compensate for this by disentangling the potential semantic relationship between texts. For training these classifiers, we use grid search algorithm to determine the best hyper-parameters for the models. Specifically, different combinations of hyper-parameters are evaluated using cross-validation [34], and the resulting optimal combination is used to build prediction models. Specifically, the hyper-parameters involved in the LR model include $\{\text{penalty}, C, \text{class weight}\}$, where penalty is valued $L1$ or $L2$ norm, C refers to the regularization parameter varying among [0.01,0.1,1.0,10,100], and class weight refers to the weights of each category, valued none or balanced; the hyper-parameters of the KNN model include $\{K, \text{weights}, p\}$, where K refers to the number of selected neighbor points, and its range of values is 1~9. Weights indicates whether the weight of the distance is taken into account in the nearest k points, valued uniform or distance, and p refers to the type of distance chosen, Manhattan Distance, Euclidean Distance or Minkowski Distance, which is meaningful only if weights = 'distance'; the hyper-parameters of the DT model include $\{\text{criterion}, \text{maximum depth}\}$, where criterion refers to the feature splitting basis metrics, gini or entory, and the maximum depth refers to the maximum depth of the tree, varying among 1–10. The hyper-parameters of the SVM model include $\{\text{kernel}, C, \text{gamma}\}$, where kernel s pecifies the kernel type to be used in the algorithm, linear function, polynomial function or radial basis function, C refers to the regularization parameter varying among [0.01, 0.1, 1.0, 10.0], and Gamma refers to the kernel coefficient valued [0.001,0.01,0.1,1.0] when the kernel type is set to 'radial basis function'.

2.5. Evaluation metrics

Five metrics, precision (Prec), recall (Rec), F1 value, accuracy (Acc), and area under the receiver operating curve (AUC), are used to measure the drug efficacy prediction performance for each model in experiments. In particular, AUC is defined as the area under the receiver operating curve, and the larger the AUC value, the better the classification effect [35]. The other metrics can be calculated as follows:

$$Prec=\frac{Tp}{Tp+Fp}$$

$$Rec=\frac{Tp}{Tp+Fn}$$

$$F1 = 2\frac{Prec \cdot Rec}{Prec + Rec}$$

$$Acc = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$$

where T_p and T_n represent the numbers of true positives and true negatives that the model correctly predicts, respectively, F_p and F_n represent the numbers of false positives and false negatives that the model incorrectly predicts, respectively.

3. Experimental results

3.1. *Lung cancer clinical data collection*

We collected the medical records of 1859 patients (2016–2020) with lung cancer from the Neusoft EMR system of our collaborated hospital locally in adherence with the Declaration of Helsinki and its later amendments. Patient inclusion criteria were: (i) patients with NSCLC and who had received first-line platinum-based chemotherapy; (ii) patients with treatment outcome available after one or two courses of treatment.

Our aim is to predict the efficacy of platinum chemotherapy based on the text data reported by radiology examination. According to the RECIST criteria, drug efficacy can be complete remission (CR), partial remission (PR), stable disease (SD) or disease progression (PD). For simplicity, we consider the binary classification problem of drug efficacy, *i.e.*, responsive (1) and non-responsive (0). The former includes CR, PR, SD, and the latter only PD. After removing the sample of patients missing clinical imaging examination, we finally made a clinical text-drug efficacy data set consisting of 958 NSCLC patient samples, of which 691 were responsive and 267 were non-responsive to platinum chemotherapy, for use of method evaluation. The clinical data of the patient population are provided in Supplementary Table S1 for reference.

We then preprocessed the raw radiology examination reports text for obtaining the word vectors for each patient. The text preprocessing includes the following three steps: 1) deleting numbers and letters; 2) segmenting words and 3) removing stop words. First, numbers and special letters can not be addressed by text analysis methods, and need to be deleted before further analysis. Segmenting words means cutting a continuous sentence into individual words. Accurate word segmentation can greatly improve the computer's ability to understand text information [36]. Stop words include prepositions, quantifiers and some non-medical nouns *etc*., which are often meaningless to text understanding and need to be removed after word segmentation. There are two word segmentation algorithms, NLPIR [37] and Jieba [38], that can be used in Chinese word segmentation. Compared with NLPIR, Jieba can tag parts of speech, and combine on the results from dictionary and statistics. We tried the two segmentation methods in the study, as shown in Fig. 2A-B, and found that Jieba retained more clinical features of patients and lesion conditions and led to better prediction results (Fig. 2 and Supplementary Fig.S1). Accordingly, we used the result of Jieba for subsequent analysis. In summary, all text documents of 958 patient samples were divided into 2372 words, most of which characterize the morphological and pathological changes of the tumor and the lesion condition that are potentially associated with treatment outcome, such as "unseen, echogenicity, expand". Totally, there are 62 terms related to lung and one parenchyma term, which would help guide in therapy of lung cancer. Details of word segmentation are given in Supplementary Table S2.

3.2. *Clinical text encoding results based on LDA model*

To perform the LDA-based clinical text encoding, we first determined the optimal number of hidden topics behind the data by examining the perplexity. Fig. 3 (A) plots the perplexity changing curve with topic number. From this figure, we can clearly see that the perplexity reaches the lowest at 16, indicating the optimal number of topics 16. Then, with $K = 16$ and $\alpha = \beta = 0.0625$ as default, we learned the LDA topic model on the data set and generated the topic distribution of each patient document, *i.e.* the encoding vector and the word distribution of each topic.

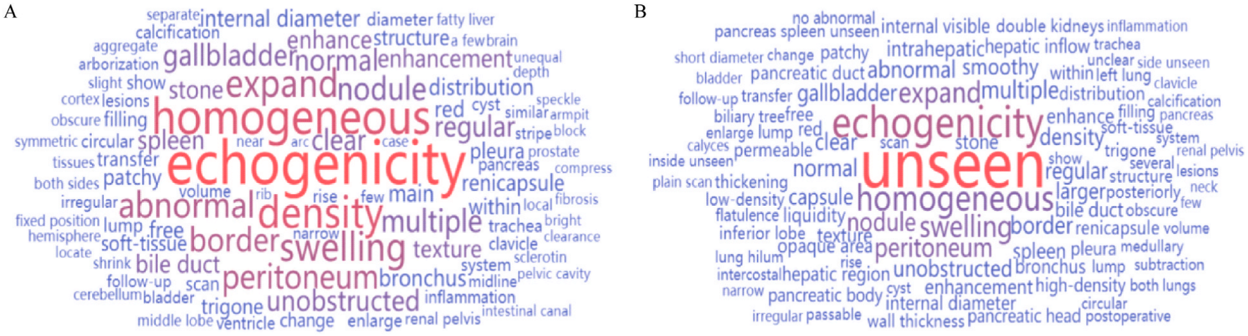

Fig. 2. Word cloud maps generated by NLPIR (A) and Jieba (B).

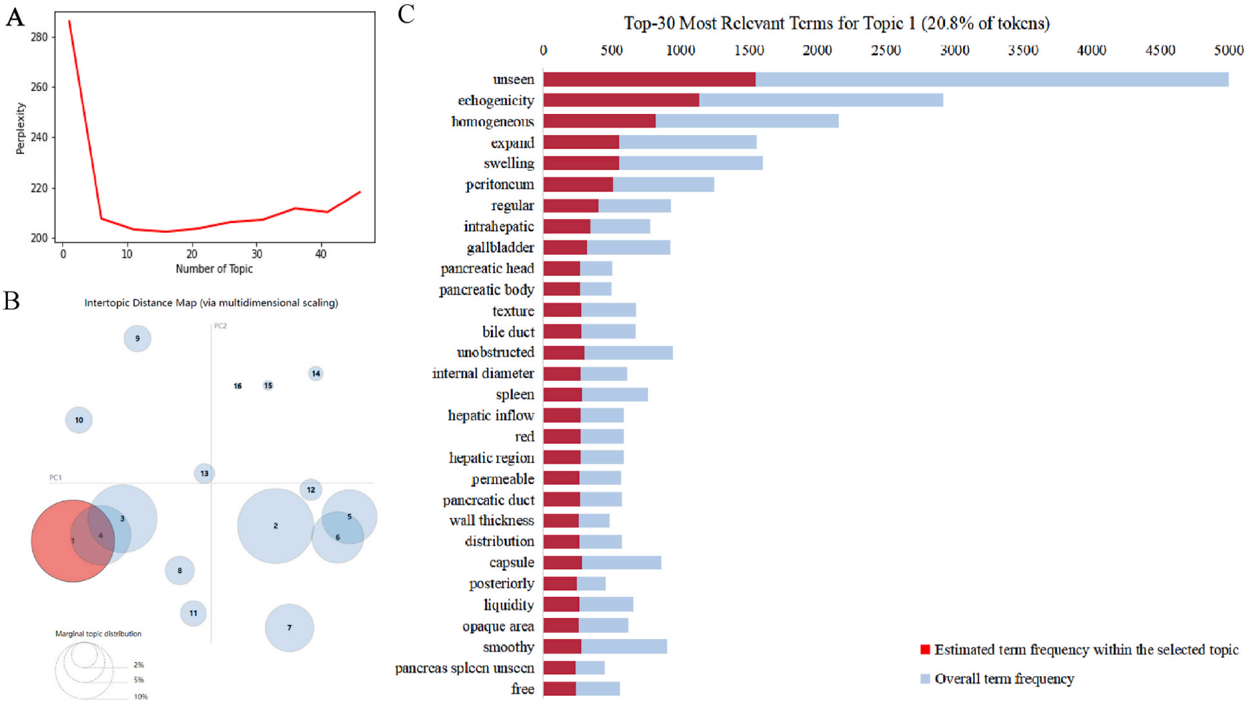

Fig. 3. LDA encoding of clinical text data. (A) Perplexity-topic number curve, indicating the perplexity arrives least at 16 topics. (B) Global view of the 16 topics in the first two PCs space by multidimensional scaling analysis, suggesting that the 16 topics are non-redundant and proper. The size of the circles indicates the prevalence of the topic, *i.e.* the frequency of topics appearing in the document corpus. (C) The frequencies of the top-30 most relevant terms for Topic 1, showing the primary word distribution pattern of the topic.

In order to analyze the result, we used multi-dimensional scaling algorithm, implemented in the pyLDAvis package [39], to visualize the obtained topics as shown in Fig. 3 (B). From this figure, it can be found that the topics scatter the whole space and most of them distribute in the negative direction of the 2nd PC. It is also noticed that 7/10 topics have >5% prevalence, of which most prevalent Topic 1 has a prevalence of >10%. As shown in Fig. 3 (C), Topic 1 has a very similar word distribution to the overall, such as the words, “unseen”, “echogenicity”, “homogeneous”, appears most frequently both in the topic and in the corpus. The highest-frequency words under this topic are mostly descriptive of patients’ stable conditions. Mining and analysis of the high-frequency words in each topic can help find text patterns associated with treatment outcome. In what follows, we will explore the encoding results to find potential associations between topic features and efficacy.

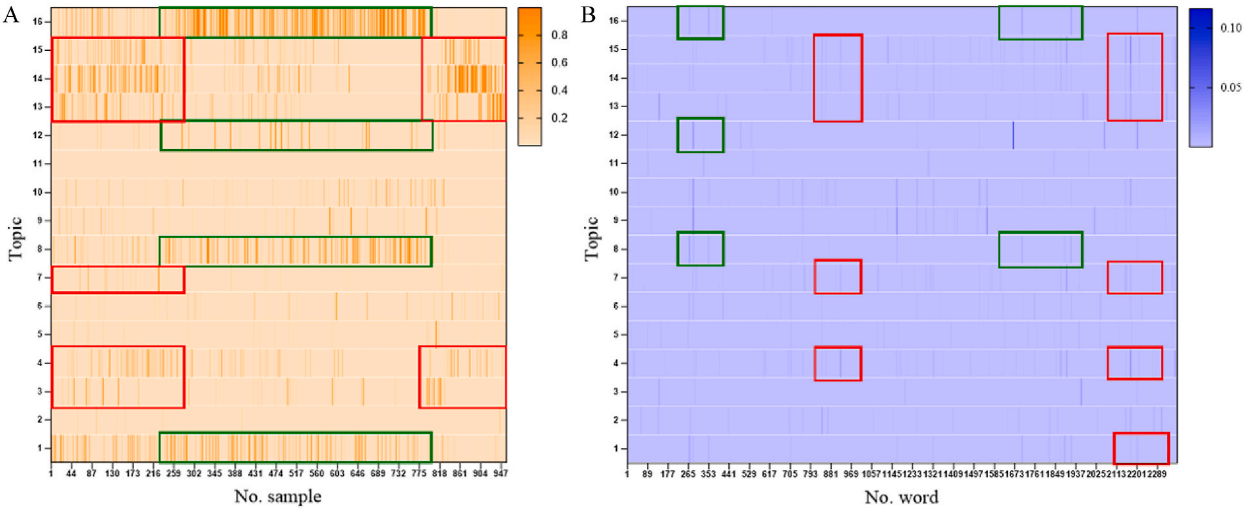

Fig. 4. Heatmaps of the topic-sample distribution matrix (A) and topic-word distribution matrix (B). We can see that some topics show similar patterns between samples/words, e.g. those within the boxes.

Fig. 4 shows the heatmaps of the obtained topic-sample distribution matrix and topic-word distribution matrix. FromFig.4 A, it can be seen that topics 1, 3, 4, 7, 8, 12, 13, 14, 15, 16 have higher probabilities in most samples, and seem to be divided into two topic groups with different distribution patterns: one group includes 1, 8, 12, 16, and another includes 3, 4, 7, 13, 14, 15. Similar results can be observed in the word-topic distribution heatmap (Fig. 4B). These results show that the LDA model can clearly mine the potential topic features underlying the clinic text data.

We further calculated the mean vectors of topics for the two sample groups: responsive and non-responsive, and compared them in each topic channel, as shown in Fig. 5A–B. We found that the responsive group took significantly higher values than the non-responsive group on feature topics 8, 12, 16, while the non-responsive group took significantly higher values than the responsive group on feature topics 4, 7, 13, 14, 15. These are in accordance with the analyses above. Accordingly, we can divide the whole 16 topics into three topic groups, drug responsive-related, drug non-responsive-related and neutral ones, of which the drug responsive-related group (RG) consists of topics 8, 12, 16, and the drug non-responsive-related group (NG) consists of topics 4, 7, 13, 14 and 15.

Fig. 6 (A-C) compares the word frequency spectra of topics among the three topic groups. From this figure, it can be clearly seen that for the RG and NG groups, the intra-group topics take on more similar spectrum patterns but the inter-group topics more different patterns. Deep investigation indicates that the highest-frequency words in the topic group RG include many seemly good condition words, such as “unseen, homogeneous, normal”, while the highest-frequency words in the topic group NG are descriptive of sound bad conditions, such as “multiple, density, nodule”, as shown in Fig. 6 (D-E). Note that the term “unseen” are often followed by ‘abnormal signal’, ‘abnormal enhancement lesion’, ‘abnormal density shadow’, etc, which mostly indicate no obvious abnormality found in the radiomic images. More biological and medical explanations are needed in future.

3.3. Performance validation of drug efficacy prediction models

We next evaluated the drug efficacy prediction performance of the LDA encoding representation. To avoid randomness, we used stratified five-fold cross validation to test the prediction performance. The main idea is to use different training/test splits for pertinent evaluation. Specifically, the whole dataset S was divided into 5 folds (disjoint subsets), *i.e.* S_i, with $i = 1,2, \ldots,5$, where the proportion of each category in each fold is the same as that in the whole dataset S. For each fold S_i, a classification model is trained on all the left folds and tested on S_i [40]. The multiple training/test splitting makes the evaluation more objective and more reliable. Note that, considering fewer non-responsive samples in the training set, we used Smote algorithm [41] to expand the non-responsive class to have the same (521) samples as the responsive samples for sample balance during training. The Smote algorithm, as a sampling technique, can synthesize new samples independently identically distributed by simulation synthesis. Specifically, it first randomly selects a sample from the minority class and calculates the Euclidean distance between it and other samples in the class to find its neighboring samples, and then synthesizes new samples by linear random interpolation. Compared with random sampling, the Smote algorithm can avoid overfitting caused by duplicated samples. Fig. 7 shows the sample distributions of training set before (Fig. 7A) and after (Fig. 7B) applying Smote algorithm, indicating that the distribution of expanded samples is the same as or similar to the original one for the class while balancing the two classes. Note that the visualization is based on the two components obtained using t-distributed Stochastic Neighbor Embedding (*t*-SNE) [42].

We then trained four classifiers, namely LR, KNN, DT and SVMs, on the expanded training sets. During training, grid search algorithm was used to find the optimal model parameters for each classifier based on internal five-fold cross validation (CV) on the training set. Table 1 shows the five-fold CV results. From this table, we can clearly see that all the four classifiers in combination with LDA encoding resulted in the AUCs of higher than 0.7, of which SVMs achieved the best drug efficacy prediction performance: a precision of 0.91, an F1 value of 0.81, an accuracy of 0.75 and an AUC of 0.77, suggesting the effectiveness of the proposed method. We also examined the computing cost of the proposed method. Running on Windows 10 (64-bit) OS with Intel(R) Core (TM) i7-7500 CPU @ 2.50 GHz, 8G RAM, under python 3.8 software, it took 102.52 s for training and 0.82 s for testing on a new sample, which are acceptable in real application situations.

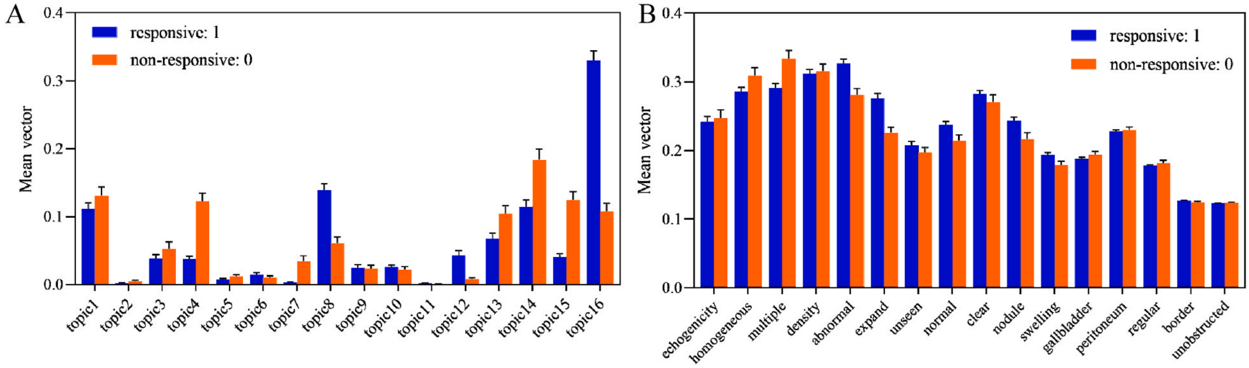

Fig. 5. Distributions of the mean vectors of the two sample classes on 16 LDA topics (A) and on the top-16 scored terms by TF-IDF model (B). We can see that the two classes show more differences on the LDA topics than on the top TF-IDF features.

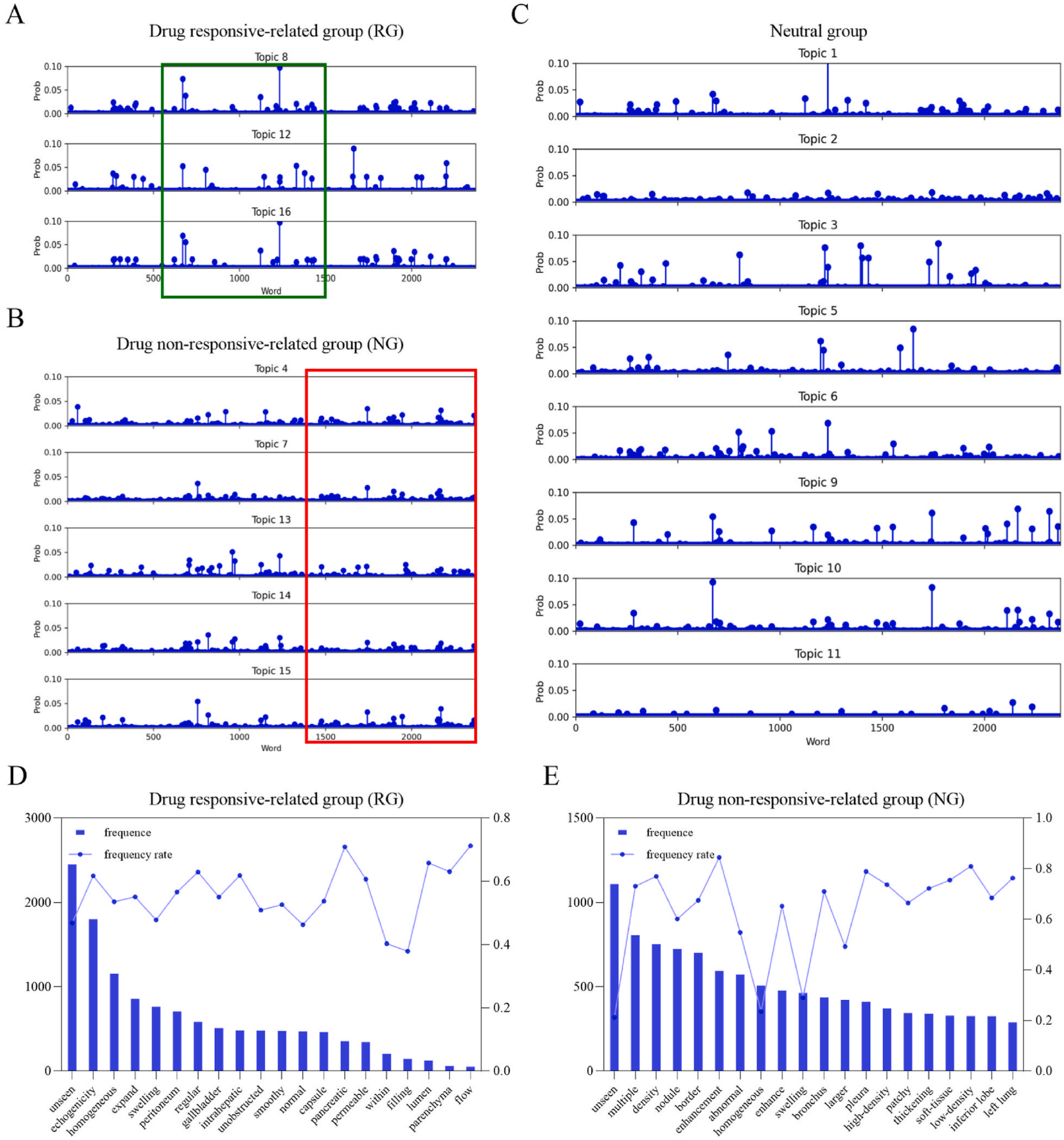

Fig. 6. Word distributions (A–C) of each topic in the three topics groups and Top-20 high-frequency words (D–E) in the responsive-related and non-responsive-related topics. We can see that each topic group show similar changing patterns over words, e.g. those within the boxes.

3.4. Statistical significance analysis

Based on the LDA + SVM model, we further verified the statistical significance of the drug efficacy prediction by permutation test. In the permutation test, we randomly shuffled the labels of training samples in the five-fold CV procedure to retrain the prediction model and re-predicted the test set. The permutation process was repeated 1000 times, and the proportion of AUC values greater than the real case (0.77) was counted as a p-value. Finally, we obtained a p-value $= 0.00 < 0.05$, indicating the statistical significance of the observed results at an *ad hoc* p-value cutoff of 0.05. Fig. 8 compares the real ROC curve with five permuted ones after randomly shuffling the labels of training samples. From this figure, it can be seen that the random AUC values are all around 0.5 as expected, which is much smaller than the real AUC value. These results show that the LDA-based drug efficacy prediction is statistically

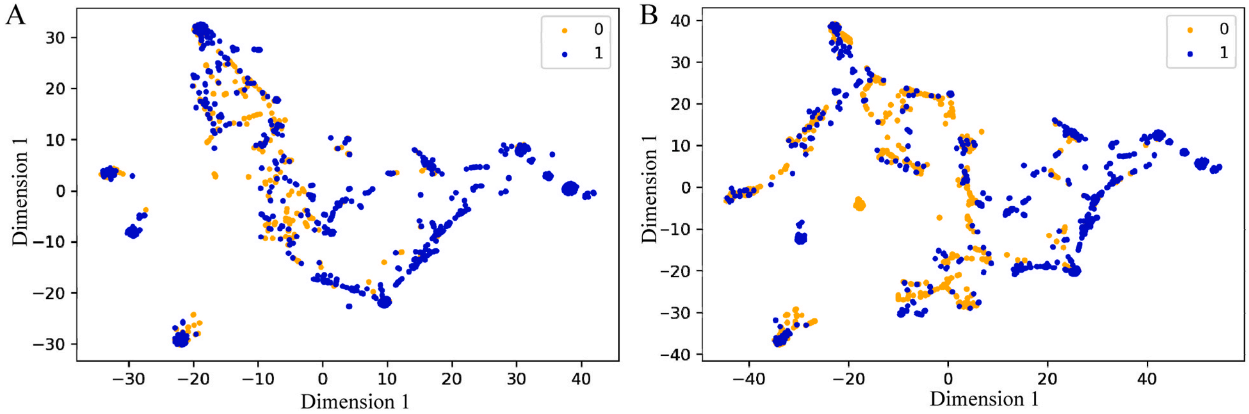

Fig. 7. t-SNE projection visualization of the training set before (A) and after (B) applying the Smote algorithm. The expanded distribution is the same as or similar to the original one while balancing the two classes.

Table 1
Drug efficacy prediction performance of the two text encoding models, LDA and TF-IDF, in combination with four classifiers (mean ± sd). Note: AUC means the area under the receiver operating curve.

Classifiers	Text Encoding models	Precision	Recall	F1	Accuracy	AUC
Logistic Regression	TF-IDF	0.83 ± 0.02	0.72 ± 0.06	0.77 ± 0.04	0.69 ± 0.04	0.66 ± 0.03
	LDA	0.89 ± 0.02	0.65 ± 0.03	0.75 ± 0.02	0.69 ± 0.03	0.72 ± 0.03
k-Nearest Neighbor	TF-IDF	0.82 ± 0.02	0.73 ± 0.03	0.77 ± 0.01	0.69 ± 0.01	0.66 ± 0.02
	LDA	0.85 ± 0.01	**0.77** ± **0.06**	0.81 ± 0.04	0.73 ± 0.04	0.7 ± 0.03
Decision Tree	TF-IDF	0.82 ± 0.02	0.75 ± 0.06	0.78 ± 0.03	0.7 ± 0.03	0.67 ± 0.03
	LDA	0.86 ± 0.02	0.74 ± 0.05	0.8 ± 0.03	0.73 ± 0.03	0.71 ± 0.02
Support Vector Machines	TF-IDF	0.85 ± 0.02	0.72 ± 0.04	0.78 ± 0.03	0.7 ± 0.03	0.7 ± 0.03
	LDA	**0.91** ± **0.02**	0.73 ± 0.03	**0.81** ± **0.01**	**0.75** ± **0.01**	**0.77** ± **0.02**

significant and justify the utility of clinical text data in drug efficacy prediction.

3.5. Comparison with TF-IDF text encoding model

For comparison evaluation, we applied the conventional Term Frequency-Inverse Document Frequency (TF-IDF) model [43] to alternatively encode the patient text document. The essential meaning of TF-IDF algorithm is: if a word appears more frequently in a document and less frequently in other documents, the word is considered to be able to differentiate effectively between documents. Briefly speaking, TF-IDF algorithm explicitly extracts text features by calculating TF-IDF value of feature words, which is obtained by multiplying TF value and IDF value. By restricting the thresholds of TF and IDF, the TF-IDF model can effectively filter common high-frequency words and retain the relatively important feature words with distinguishing ability in the document.

For fairness, we selected the top 16 word terms with highest TF-IDF values, and calculated the center vectors of the two classes on the 16 word terms, as shown in the right side of Fig. 5. From this figure, it can be found that the center vectors of the two classed have less differences compared with those by the LDA model (the left subfigure of Fig. 5). Then, we applied the four classifiers to the TF-IDF text encodings of patients and compared the resulting efficacy prediction performances with those by the LDA model, as shown in Table 1. From this table, we can clearly see that the LDA model outperforms TF-IDF, irrespective of the classifier used, suggesting the superior performance of LDA in extracting drug efficacy-associated information from the clinical text data. The TF-IDF algorithm is based on the bag-of-words idea and only extracts features based on “word frequency”. In contrast, the LDA model introduces the hidden variable of topic between words and documents, which can deal well with “Polysemy” and “More word one meaning”. By mining deeper information in the semantic level, LDA can discover the association patterns between drug efficacy and patient condition underlying the texts.

3.6. Comparison with clinical structured data

Structured data are another kind of rich information in clinical data, possibly useful to drug efficacy prediction. There are totally 247 structured patient features in the lung cancer data, *e.g.*, biometric information (gender, age), vital signs, history of past illness, laboratory tests. We preprocessed the structured data of the 958 patients by the following steps: removing patient features with missing values greater than 50% and for the remaining, missing values were imputed using the mean values. Finally, 56 structured feature variables were obtained and normalized to a variance of 1 and a mean of 0 for further analysis. Details about the structured data are given in Supplementary Table S3.

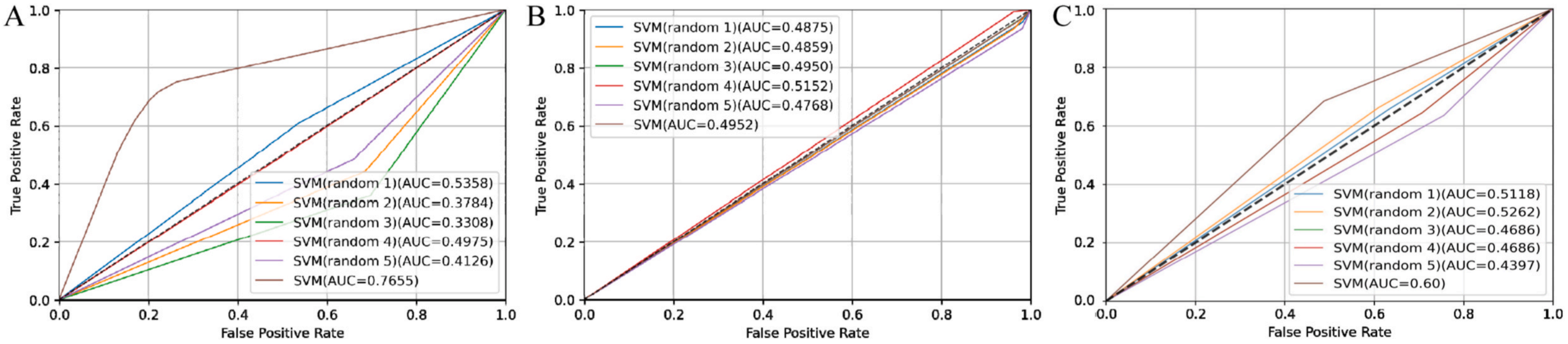

Fig. 8. Real and random ROC curves on the clinical text data (A) and clinical structured data (B) of lung cancer patients and on the clinical text data of bowel cancer patients (C). We can see that significant prediction can be obtained on the clinical text data rather than the clinical structured data. Note: SVM means support vector machine; AUC means the area under the receiver operating curve.

We first combined the structural data with the text data above to represent each patient for drug efficacy prediction. Table 2 lists the five-fold CV results by the four classifiers, LR, KNN, DT and SVMs, on the combined data. As can be seen from this table, there is no increase, but even a decrease, in almost all evaluation metrics after adding the structured data. To account for this, we next predicted drug efficacy using the structured data alone. As a result, the four classifiers obtained AUCs of 0.51, 0.51, 0.5 and 0.5, respectively, which all are far smaller than those by the clinical test data (Table 1), as shown in Table 2. These may suggest that the structural data may be not predictive of drug efficacy. Furthermore, we estimated the statistical significance of the results in a permutation test similar to the above. For the real AUCs by LR, KNN, DT and SVMs, the resulting p-values are far larger than an *ad hoc* cutoff of 0.05 (0.31, 0.36, 0.4 and 0.35, respectively), meaning non-significant predictions, as shown in Fig. 8 (right). This may be related to two things: too many missing values and inclusion of too few structured data, which needs to be further verified.

3.7. Validation on an independent dataset

To further evaluate the proposed method, we collected 266 bowel cancer patients (2020–2022) treated with platinum from The First Affiliated Hospital of University of Science and Technology of China (ethical approval no. 2021- RE-85) as an independent data set. Similar to the lung cancer data, the efficacy of cisplatin in these patients was also divided into two categories: responsive (1) and non-responsive (0) for binary classification, resulting in 225 responsive patients and 41 non-responsive patients. We applied the trained LDA + SVM model to predict the efficacy categories of the bowel cancer patients. The results are as follows: a precision of 0.89 a recall rate of 0.68, an F1 value of 0.77, an accuracy of 0.66, and an AUC value of 0.6, which, albeit slightly lower than those on the lung cancer data, indicating the prediction effectiveness on the independent data set. To examine the statistical significance of the prediction, similar to the permutation test in Section 3.4, we randomly shuffled the labels of the 266 bowel cancer patients 1000 times, and calculated the p-value of the observed AUC result. As a result, a $p\text{-value} = 0.00 < 0.05$ were obtained, indicating that the observed result is statistically significant at an *ad hoc* p-value cutoff of 0.05. Fig. 8 (right) illustrates the observed and five randomized ROC curves. These results suggest the prediction generalizability of the proposed method on independent data sets.

4. Conclusion

In this paper, we have proposed to predict drug efficacy based on clinical text data for cancer patients. The proposed method used LDA models to encode imaging examination text, and based on the text encoding features, ML classifiers were then employed for drug efficacy prediction. LDA models can mine the hidden topics underlying the clinical text data that reflect the patterns of tumor tissue characteristics associated with treatment outcome, which benefits the prediction capability of the proposed method. In experiments, we established clinical text data sets including lung and bowel cancer patients treated with platinum to evaluate the proposed method. The experimental results on the data sets show the efficacy and effectiveness of the proposed method, and especially, demonstrate that the clinical text data are predictive of personalized drug efficacy for individual cancer patients, suggesting the potential clinical application value.

Currently, cancer patients are treated empirically in clinic, which has a non-response rate of up to 30–40%, for example, to cisplatin. Realizing cancer precision medicine needs to overcome the challenging individual heterogeneity of treatment. Compared with genomic data-based methods, those based on clinical data easily available, e.g., radiology examination reports, are cost-effective and may alternatively provide a new promising paradigm for solving the challenge. On the other hand, we also notice that the proposed method still has some limitations, e.g. only using a few types of clinical data and segmenting words only in unigram mode. Future work will be focused on introducing more types of clinical data, such as histological reports, as well as improving the text processing and classification models for better drug efficacy prediction.

Funding

This work was supported in part by the National Natural Science Foundation of China [grant numbers 61973295, 81872276]; Key Research and Development Project of Anhui Province (grant number 201904a07020092); University Science Research Project of the Education Department of Anhui Province (grant number KJ2021A0633).

Data availability statement

Data and code will be available on our server http://aisys.iim.ac.cn/download.html and be free for non-profit use.

Ethics statement

This study was reviewed and approved by the Ethics Committee of The First Affiliated Hospital of University of Science and Technology of China, with the approval number: 2021-RE-85. Informed consent was not required for this study because the data are anonymized.

CRediT authorship contribution statement

Xinping Xie: Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis,

Table 2
Prediction results using structured data and combined data (mean ± sd). Note: AUC means the area under the receiver operating curve.

Data types	Classifiers	Precision	Recall	F1	Accuracy	AUC
Structured	Logistic Regression	0.73 ± 0.02	0.6 ± 0.05	0.66 ± 0.03	0.55 ± 0.03	0.51 ± 0.03
	k-Nearest neighbor	0.73 ± 0.02	0.72 ± 0.03	0.72 ± 0.02	0.6 ± 0.03	0.51 ± 0.04
	Decision Tree	0.72 ± 0.02	0.7 ± 0.1	0.71 ± 0.04	0.59 ± 0.02	0.5 ± 0.03
	Support vector machines	0.72 ± 0.00	0.95 ± 0.01	0.82 ± 0.01	0.7 ± 0.01	0.5 ± 0.01
Text + Structured	Logistic Regression	0.79 ± 0.01	0.69 ± 0.05	0.73 ± 0.03	0.64 ± 0.02	0.61 ± 0.01
	k-Nearest neighbor	0.73 ± 0.02	0.66 ± 0.04	0.69 ± 0.02	0.58 ± 0.02	0.52 ± 0.03
	Decision Tree	0.83 ± 0.02	0.78 ± 0.04	0.81 ± 0.01	0.73 ± 0.01	0.69 ± 0.02
	Support vector machines	0.72 ± 0.00	0.96 ± 0.01	0.82 ± 0.01	0.7 ± 0.01	0.5 ± 0.01

Conceptualization. **Dandan Li:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Yangyang Pei:** Writing – original draft, Visualization, Formal analysis, Data curation. **Weiwei Zhu:** Writing – original draft, Investigation, Formal analysis. **Xiaodong Du:** Validation, Investigation, Formal analysis. **Xiaodong Jiang:** Visualization, Validation, Resources, Investigation, Data curation. **Lei Zhang:** Validation, Resources, Data curation. **Hong-Qiang Wang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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.

Abbreviations

Electronic medical records (EMRs)
Cancer Cell Line Encyclopedia (CCLE)
Genomics of Drug Sensitivity in Cancer (GDSC)
Support Vector Machines (SVMs)
Convolutional neural network (CNN)
Half maximal inhibitory concentration (IC50)
Computed tomography (CT)
B-mode ultrasound (BU)
Magnetic Resonance Imaging (MRI)
Latent Dirichlet Allocation (LDA)
Machine learning (ML)
Logistic Regression (LR)
k-Nearest Neighbor (KNN)
Decision Tree (DT)
Term Frequency-Inverse Document Frequency (TF-IDF)
Area under the receiver operating curve (AUC)
t-distributed Stochastic Neighbor Embedding (*t*-SNE)
Cross validation (CV)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e27300.

References

[1] T. Sakellaropoulos, et al., A deep learning framework for predicting response to therapy in cancer, Cell Rep. 29 (11) (2019) 3367–3373. e4.
[2] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 71 (3) (2021) 209–249.
[3] M. Nicholas, S. Charles, Clonal heterogeneity and tumor evolution: past, present, and the future, Cell 168 (4) (2017) 613–628.
[4] J. Barretina, et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (7391) (2012) 603–607.
[5] M.J. Garnett, et al., Systematic identification of genomic markers of drug sensitivity in cancer cells, Nature 483 (7391) (2012) 570–575.
[6] S.d. Li, y.s. Li, A computational model for predicting classification of anticancer drug response to individual tumor and its applications, Prog. Biochem. Biophys. 49 (6) (2022) 1165–1172.
[7] J. Bai, R. Han, C.a. Guo, A hybrid convolutional network for prediction of anti-cancer drug response, in: 2021 11th International Conference on Information Science and Technology, ICIST), 2021.
[8] M. Li, et al., DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines, IEEE ACM Trans. Comput. Biol. Bioinf 18 (2) (2019) 575–582.

[9] Y. Fang, et al., A quantile regression forest based method to predict drug response and assess prediction reliability, PLoS One 13 (10) (2018) e0205155.
[10] j.b. Wang, et al., Application status of bigdata in clinical treatment and its challenges to education, Education And Teaching Forum (35) (2022) 43–46.
[11] L. Dercle, et al., Radiomics response signature for identification of metastatic colorectal cancer sensitive to therapies targeting EGFR pathway, J. Natl. Cancer Inst. 112 (9) (2020) 902–912.
[12] Q. Zeng, et al., Editorial: imaging technology in oncology pharmacological research, Front. Pharmacol. 12 (2021) 711387.
[13] S. Wang, et al., *Mining whole-lung information by artificial intelligence for predicting EGFR genotype and targeted therapy response in lung cancer: a multicohort study.* The Lancet, Digital health 4 (5) (2022) e309–e319.
[14] x.l. Zhu, j.l. Wu, Application progress of MRI radiomics in the efficacy and prognosis of neoadjuvant chemotherapy for breast cancer, Chinese Journal of Magnetic Resonance Imaging 13 (3) (2022) 159–161+165.
[15] Y. Zhang, et al., Complex cystic and solid breast lesions: diagnostic performance of conventional ultrasound, strain imaging and point shear wave speed measurement, Clin. Hemorheol. Microcirc. 69 (3) (2018) 355–370.
[16] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, ACM SIGMOD Record 22 (2) (1993) 207–216.
[17] J. Macqueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.
[18] D.J. Hand, Text mining: classification, clustering, and applications edited by ashok srivastava, mehran sahami, Int. Stat. Rev. 78 (1) (2010) 134–135.
[19] B. Shen, et al., A systematic assessment of deep learning methods for drug response prediction: from in vitro to clinical applications, Briefings Bioinf. 24 (1) (2022) 1–13.
[20] A. Rumshisky, et al., Predicting early psychiatric readmission with natural language processing of narrative discharge summaries, Transl. Psychiatry 6 (10) (2016) 1–5.
[21] A. Zalewski, et al., Estimating patient's health state using latent structure inferred from clinical time series and text, IEEE-EMBS International Conference on Biomedical and Health Informatics., 2017 (2017) 449–452.
[22] M.C. Hughes, et al., Assessment of a prediction model for antidepressant treatment stability using supervised topic models, JAMA Netw. Open 3 (5) (2020) e205308.
[23] s. Liang, Design and Implementation of a Structured Processing System for Pathological Text Data, Donghua University, 2015.
[24] C. Meilan, et al., Analysis of the impact of investor sentiment on stock price using the latent dirichlet allocation topic model, Front. Environ. Sci. 10 (2022) 106939.
[25] J. Ma, et al., An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local, Expert Syst. Appl. (2023) 212.
[26] X. Liu, et al., Intelligent radar software defect classification approach based on the latent Dirichlet allocation topic model, EURASIP J. Appl. Signal Process. (44) (2021).
[27] L.d. Wang, et al., Prescription function prediction using topic model and multilabel classifiers, Evid. Base Compl. Alternative Med. (2017) 8279109.
[28] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[29] y.w. Liu, y. Zhang, s. Yang, Disease-assisted diagnosis based on LDA model and electronic medical record, Journal of Suzhou University 32 (2) (2017) 114–116+124.
[30] M.J. Wainwright, M.I. Jordan, Graphical models, exponential families, and variational inference, Foundations and Trends in Machine Learning 1 (1–2) (2008) 1–305.
[31] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. 39 (1) (1977) 1–38.
[32] j. Wu, et al., LDA feature selection based text classification and user clustering in Chinese online health community, Journal of the China Society for Scientific and Technical Information 36 (11) (2017) 1183–1191.
[33] S. Garg, Drug recommendation system based on sentiment analysis of drug reviews using machine learning, in: 11th International Conference on Cloud Computing, Data Science and Engineering, Noida, 2021.
[34] j.x. Liu, Support vector regression based on grid search hyperparameter optimization, Scientific and Technological Innovation (13) (2022) 71–74.
[35] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (1) (2011) 37–63.
[36] Z. Wu, G. Tseng, Chinese text segmentation for text retrieval: achievements and problems, J. Am. Soc. Inf. Sci. 44 (9) (1993) pp532–542.
[37] H.p. Zhang, J.y. Shang, NLPIR-Parser: an intelligent semantic analysis toolkit for big data, Corpus Linguistics 6 (1) (2019) 87–104.
[38] x.q. Zeng, Technology implementation of Chinese Jieba segmentation based on Python, China Computer & Communication 31 (18) (2019) 38–39+42.
[39] C. Sievert, K.E. Shirley, LDAvis: a method for visualizing and interpreting topics, in: Workshop on Interactive Language Learning, visualization and interfaces, Baltimore, 2014.
[40] R. Bey, et al., Fold-stratified cross-validation for unbiased and privacy-preserving federated learning, J. Am. Med. Inf. Assoc. 27 (8) (2020) 1244–1251.
[41] N.V. Chawla, et al., SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[42] V.D.M. Laurens, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2605) (2008) 2579–2605.
[43] S. Gerard, Developments in automatic text retrieval, Science 253 (5023) (1991) 974–980.