Postprint of a Study on Control Mechanisms and Models for Overuse and Inappropriate Medical Testing in Oncological Diseases

Authors: Zhu Shisheng, Wang Xinrong, Mao Liting, Xueguo Liu

Date: 2018-07-23T00:00:00+00:00

Abstract

To address the existing problems of errors and over-medicalization in the current diagnosis and treatment of neoplastic diseases, this paper proposes an inspection and control mechanism and solution for identifying erroneous and excessive medical interventions by extracting imaging information from expert prescriptions of similar cases based on medical big data and utilizing machine learning classification models. This solution leverages CT and MRI images accumulated over the long term in medical records of various tumor diseases in hospitals. Using the actual tumor type in each diagnostic and treatment process as the ground truth, corresponding imaging data types are selected from the medical database for feature extraction, feature selection, and model construction to obtain a predictive classifier for that tumor type, which predicts the benign or malignant status of the current case. Issues of over-treatment and diagnostic errors in the treatment process are identified through comparison with physicians' diagnostic results. The core objective is to enhance the discrimination accuracy of automated methods to reduce the likelihood of misdiagnosis in tumor diseases. Experimental results demonstrate that SVM_RFE dimensionality reduction combined with Spearman redundancy removal achieves superior performance in SVM models for lung nodule benign-malignant classification compared to traditional SVM_RFE methods, and also outperforms conventional radiomics approaches. This solution can promptly detect errors and over-treatment issues and issue early warnings, serving a supervisory and reminder function. While preventing and avoiding diagnostic and therapeutic errors, it reduces reliance on manual identification and provides a novel approach for addressing medical errors and alleviating patient burden.

Full Text

Abstract

Aimed at addressing the pervasive issues of erroneous and excessive medical treatment in tumor disease diagnosis and management, this study proposes a control mechanism and solution for detecting misdiagnosis and overtreatment by leveraging machine learning classification models. The approach extracts imaging information from expert prescriptions in similar medical cases based on large-scale medical data. Relying on CT and MRI images accumulated in hospital databases across various tumor cases, the method selects corresponding imaging data for each tumor type encountered during clinical practice, performs feature extraction and selection, and constructs predictive classifiers to determine tumor malignancy. By comparing these predictions with physician diagnoses, the system identifies potential instances of overtreatment or misdiagnosis. The core objective is to reduce diagnostic errors in tumor diseases by improving classification accuracy independent of manual judgment. Experimental results demonstrate that SVM_RFE dimensionality reduction combined with Spearman redundancy elimination outperforms traditional SVM_RFE methods in SVM-based classification of pulmonary nodule malignancy, while also surpassing conventional radiomics approaches. This solution enables timely detection and early warning of medical errors and excessive treatment, serving a supervisory and reminder function. It reduces reliance on manual differentiation while preventing diagnostic mistakes, offering a novel pathway to mitigate erroneous medical practices and alleviate patient burden.

Keywords: erroneous medical treatment; machine learning; Spearman; SVM_RFE; SVM classifier

0 Introduction

According to "Cancer Statistics in China, 2015" published by Professor Chen Wanqing and colleagues from the National Cancer Center in January 2016, China—a country with 1.37 billion people—saw an estimated 4.292 million new cancer cases and 2.814 million cancer deaths in 2015 alone. Lung cancer incidence and mortality rank highest among all cancers, with one Chinese person dying from lung cancer every 1.2 minutes on average [1]. While scientific progress has expanded treatment options, the proliferation of drugs and novel therapies has exacerbated problems of medical errors and overtreatment. Erroneous and excessive medical treatment refers to situations where physician subjectivity or negligence worsens patient conditions or increases patient burden during diagnosis and treatment, including but not limited to misdiagnosis, medication errors, and surgical mistakes [2].

Tumor diseases carry high mortality rates and treatment costs, and misdiagnosis inflicts tremendous suffering and financial burden on patients. Current clinical diagnosis of tumor malignancy primarily relies on physicians' subjective experi-

ence combined with certain clinical features. Further confirmation occurs only after physicians suspect malignancy. For patients with benign tumors, misclassification as malignant leads to unnecessary follow-up examinations and treatments—constituting overtreatment. Conversely, malignant tumors misdiagnosed as benign cause patients to miss critical early treatment windows. Therefore, solving the benign-malignant classification problem is key to reducing erroneous and excessive medical treatment in tumor diseases.

Due to its sporadic nature and ambiguous evaluation criteria, research on medical errors has seen limited breakthrough progress. Some countries primarily rely on moral constraints and public pressure to reduce such incidents, or standardize medication procedures under professional supervision, without achieving technical monitoring and management of errors and overtreatment [3]. In 2012, Professor Lambin' s radiomics concept sparked enthusiasm among medical researchers for disease prediction based on medical imaging. Subsequent years have seen continuous advancement in research using image analysis combined with statistical linear regression methods to predict disease severity for major tumor conditions, aiming to transform the traditional practice of relying solely on physicians' film-reading experience [4]. In 2016, Carneiro demonstrated the potential of machine learning-based nonlinear classification methods for predicting 5-year mortality in elderly patients using CT images [5], offering an alternative approach to addressing medical error problems.

This study focuses on constructing highly reliable benign-malignant classification models based on tumor patient records to reduce the probability of overtreatment and misdiagnosis. For 194 collected pulmonary nodule cases, we first applied the mainstream radiomics method: extracting texture features from medical images, performing dimensionality reduction via lasso regression, modeling with logistic regression, and evaluating the model using ROC curves. We then employed alternative statistical and machine learning methods in the dimensionality reduction and modeling stages to compare different feature selection and modeling approaches. Results demonstrate that the model combining Spearman redundancy elimination with SVM_RFE dimensionality reduction and SVM classification achieves optimal performance in distinguishing benign from malignant pulmonary nodules.

1 Framework for Detecting Overtreatment and Misdiagnosis

Based on extensive tumor disease data and thorough investigation of hospital workflows, this study proposes a control mechanism and methodology for detecting erroneous and excessive treatment in tumor diseases. The mechanism begins with patient medical records, using imaging information to predict tumor malignancy through machine learning methods. By comparing these predictions with actual diagnoses, the system generates evaluation reports to provide early warning and supervision of potential misdiagnosis and overtreatment. Cases showing large discrepancies between predicted and actual diagnoses, with error

rates exceeding predefined thresholds, are submitted to higher-level regulatory bodies to help physicians identify potential issues promptly and prevent their occurrence. This solution aims to establish a scientific and standardized evaluation mechanism for overtreatment and misdiagnosis across numerous tumor conditions through radiomics-based data mining, thereby improving objectivity and reducing manual assessment dependency. The specific monitoring framework is illustrated in Figure 1 [Figure 1: see original paper].

As shown in Figure 1, the framework separates the diagnosis process (which involves high physician participation) from the evaluation process (which assesses diagnostic results), minimizing human influence on evaluation reports to enhance objectivity. By leveraging large-scale imaging data from existing medical records, the framework improves the repeatability and credibility of predictive models. When tumor patients visit hospitals, physicians combine patient symptom descriptions with medical examinations to formulate diagnoses based on personal knowledge and experience, generating medical records containing CT/MRI imaging data.

These records serve as inputs to a filter that screens and extracts features from diverse tumor imaging data in the medical database for subsequent machine learning, establishing predictive models corresponding to actual diagnostic outcomes. The imaging data from records are then fed into the trained predictor to extract discrete or continuous values from multiple feature attributes. Classification results are compared with clinical malignancy indicators (actual diagnoses) to determine whether medical errors occurred during treatment. Discrepancies between predictions and actual diagnoses flag potentially erroneous processes, triggering warnings in evaluation reports. Matching results indicate absence of errors or overtreatment, prompting the framework to consider the process compliant and add the record' s imaging data and diagnosis to the trusted medical database as new training samples, preserving both the imaging data and clinical malignancy labels for future modeling.

The medical database forms the foundation of predictive modeling, storing not only pathological and pharmaceutical knowledge but also expert prescription libraries. Database richness and accuracy directly affect model credibility. During initial database construction, ensuring model correctness requires starting with absolutely reliable samples. Therefore, early-stage tumor imaging data must derive from surgically resected cases with pathologically confirmed results serving as gold-standard labels for the images.

2 Modeling Methods for Overtreatment and Misdiagnosis Detection

2.1 Classification Model for Overtreatment and Misdiagnosis

The core of solving erroneous and excessive medical treatment lies in establishing reliable classification models to identify diagnostic errors. Support Vector

Machine (SVM) is a binary classification model defined as a linear classifier with maximum margin in feature space. Its learning strategy of margin maximization transforms into a convex quadratic programming problem, perfectly matching the tumor benign-malignant classification needs for medical error detection. This study applies SVM theory to address medical error problems.

According to SVM principles described in literature [7], given benign and malignant samples of a specific tumor disease as the training set D = {(x , y), (x , y), ···, (x , y)}, where x represents various image feature parameters extracted from medical record imaging data and y denotes the patient' s true tumor malignancy level, the fundamental classification idea involves finding a separating hyperplane in sample space that divides different categories. Let y represent the target disease category and y ···y denote other disease types uniformly labeled as y . The ideal goal is to locate a hyperplane "in the middle" between y and y training samples. In sample space, the separating hyperplane can be described by the linear equation [7]:

$$\omega^T x + b = 0$$

where is the normal vector determining hyperplane orientation, and b is the displacement term determining distance from the origin. Clearly, the hyperplane is determined by and b, denoted as (, b). The distance from any point in sample space to hyperplane (, b) can be expressed as:

$$r = \frac{|\omega^T x + b|}{\|\omega\|}$$

Assuming hyperplane (, b) correctly classifies y and y training samples, then for (x , y) D, if belonging to target disease y , we have ^T x + b 0; if not belonging to target disease y , then ^T x + b < 0, which can be written as:

$$\begin{cases} \omega^T x_i + b \geq +1, & y_i = +1 \\ \omega^T x_i + b \leq -1, & y_i = -1 \end{cases}$$

Training sample points that satisfy these conditions are called "support vectors." The sum of distances from two heterogeneous support vectors to the hyperplane is:

$$\gamma = \frac{2}{\|\omega\|}$$

This is referred to as the "margin" [7].

The workflow for the overtreatment/misdiagnosis classification model involves constructing classifiers based on tumor disease types described in medical

records and corresponding imaging data. Since CT/MRI imaging aims to correctly identify malignant tumor cases, we select a certain number of malignant case images for a specific tumor type from the medical database as the SVM training positive sample set, and an equal number of benign case images as the negative sample set. Training on these data identifies the optimal separating hyperplane, yielding the final prediction model for distinguishing tumor malignancy. The consistency between this prediction result and the physician' s diagnosis serves as the criterion for evaluating overtreatment and misdiagnosis.

2.2 Feature Selection Algorithm

Considering that major diseases yield numerous clinical features—particularly tumor diseases with imaging data where image-based feature information is vast —using SVM for predictive modeling can easily lead to overfitting due to high data dimensionality, ultimately affecting model generalization. Therefore, feature selection is crucial. For diseases with complex clinical information, feature selection reduces data volume, computational complexity, and time consumption. Moreover, not all feature attributes positively contribute to predictive models; feature selection reduces dimensionality and filters out interfering features to improve prediction accuracy.

SVM_RFE is a feature ranking and selection algorithm based on Support Vector Machines [8]. The algorithm implementation consists of three steps: (a) Train SVM on the dataset containing multiple clinical feature parameters to obtain the Lagrange multiplier vector of SVM regularization terms (classifier parameters); (b) Calculate the weight vector for each feature in input space to obtain complete SVM weights; (c) Rank weight vectors using appropriate sorting methods and remove features with minimal contribution (lowest impact on correct target disease classification) until only one feature remains.

The entire SVM weight is denoted as , while represents SVM weight after removing the p-th feature [9]. This ranking method, based on feature weights generated during SVM training, selects parameters highly correlated with classification results, achieving good prediction performance when combined with SVM classifiers. However, this sequential feature reduction approach cannot adequately consider inter-feature relationships, potentially allowing several highly correlated features to simultaneously rank at the top and compress feature selection space, preventing identification of optimal feature combinations.

To address this limitation, this study employs Spearman correlation-based redundancy elimination before SVM_RFE feature selection to remove highly correlated features.

3 Comparison of Spearman-Enhanced SVM and Traditional Radiomics Methods

3.1 Training Data Extraction Process

Following the description of SVM-based medical error modeling, the next challenge involves obtaining high-quality training data to ensure model generalization. As previously mentioned, training data originates from imaging data in existing tumor medical records within hospital databases. The specific workflow is shown in Figure 2 [Figure 2: see original paper].

Figure 2 illustrates the training data extraction process based on large-scale medical records. The pathology and pharmaceutical knowledge base stores specialized knowledge and experience from medical experts, as well as excellent medical records and prescriptions accumulated over years. Database richness and correctness directly affect final model credibility. Data appears as standardized medical records containing patient information, symptom descriptions, pathological parameters, physician diagnostic orders [10], and unstructured imaging data.

Predictive model construction extracts a certain number of malignant case images for the tumor type from the medical database as training set positive samples, and an equal number of benign case images as negative samples. Feature extraction is performed on these randomly selected training samples, followed by dimensionality reduction using the combined SVM-RFE feature selection algorithm with Spearman redundancy elimination. Therefore, standardized collection and accumulation of training data in the medical database ensures classifier prediction effectiveness.

3.2 Traditional Radiomics-Based Tumor Classification Model

As previously discussed, solving medical error problems hinges on establishing reliable classification models based on pathological indicators. Correctly classifying diseases is crucial—for tumors, effectively reducing medical errors during diagnosis and treatment significantly benefits patients by eliminating unnecessary costs while ensuring timely appropriate treatment. Current radiomics research primarily constructs tumor grading and classification models from tumor imaging data using computer vision techniques and statistical analysis/linear regression methods, separating tumor prediction from clinical experience accumulation to evaluate accuracy and AUC metrics.

Popular radiomics methods in recent years mainly consist of four steps shown in Figure 3 [Figure 3: see original paper]: (a) Data preprocessing, including CT/MRI image acquisition, region-of-interest delineation (typically lesion areas) by professional physicians, and feature extraction from grayscale images; (b) Lasso regression for dimensionality reduction on extracted features to obtain sparse solutions for linear models; (c) Logistic regression modeling for tumor benign-malignant classification based on selected features; (d) Model evalua-

tion using reserved validation data, assessing generalization capability through multiple metrics including prediction accuracy, sensitivity, and specificity [11].

3.3 Disease Prediction Model Construction Based on SVM and Similar Expert Prescriptions

The pulmonary nodule CT imaging data studied in this research is shown in Figure 4 [Figure 4: see original paper]. The left image shows that the lesion area occupies a very small proportion, while the right image displays a magnified 3D nodule view. Physicians face considerable difficulty judging nodule malignancy based solely on such imaging information, easily leading to medical errors.

Since lesions in different human tissues exhibit varying manifestations in medical images, collecting clinical features—especially imaging data—from similar cases is challenging, resulting in insufficient training sample sizes more suitable for machine learning methods with good small-sample classification performance. The modeling workflow proceeds as follows: (a) Collect N tumor CT image cases with confirmed benign-malignant pathological labels, randomly divided into training set x cases and validation set y cases; (b) Extract GLCM, GLRLM, Histogram, and Form factor texture parameters totaling M features from all samples through image recognition methods, obtaining an N×M feature matrix; (c) Perform feature engineering on training samples, combining Spearman method to screen m most effective features for classification, reducing the feature matrix to N×m (partial reduced-dimensional data shown in Figure 5 [Figure 5: see original paper]); (d) Construct SVM classification model based on x training samples and validate using y validation cases to assess disease classification prediction capability.

3.4 Comparison of Spearman-Enhanced SVM and Traditional Radiomics Methods

As described in Sections 3.1 and 3.2, both methods address tumor disease overtreatment and misdiagnosis detection identically—obtaining imaging data from databases and feeding back benign-malignant predictions. They share similar data preprocessing, feature extraction, and model evaluation stages, differing only in feature reduction and classifier construction. Radiomics builds general linear regression models (logistic regression) and employs lasso regression (L1-constrained linear regression) for sparse solutions, which inherently lacks redundancy elimination functionality. In contrast, the Spearman-enhanced SVM method uses Gaussian kernel SVM classifiers—nonlinear classifiers performing well on non-linearly separable datasets—and combines Spearman correlation-based redundancy elimination with SVM_RFE nonlinear dimensionality ranking to better preserve effective features.

4 Experimental Validation and Results

This study employs 194 real pulmonary nodule imaging cases collected from the Fifth Affiliated Hospital of Sun Yat-sen University as experimental samples. Through comparative experiments with general SVM and traditional radiomics methods, along with visualized results, the study validates the rationality and scientific validity of the proposed control mechanism and model. Further applicability research is required before clinical implementation.

4.1 Data Division and Preprocessing

The total sample consists of 194 pulmonary nodule cases (139 benign, 55 malignant). The experiment adopts a common 7:3 training-validation split ratio. Random selection yields 135 training cases (98 benign, 37 malignant) and 59 validation cases (41 benign, 18 malignant). Image processing extracts GLCM, GLRLM, Histogram, and Form factor texture features totaling 330 parameters. The training and validation datasets undergo standardization and outlier mean imputation. Malignant tumors receive label 0, benign tumors label 1. Post-division sample distributions are shown in Table 1 .

4.2 Feature Reduction and Modeling

Dimensionality reduction uses only training data without involving validation data. Validation set features corresponding to selected dimensions are extracted only during model evaluation for input to each model.

Following traditional radiomics methods, lasso algorithm reduces 330-dimensional training data (135 cases) directly to a 10-dimensional sparse solution, achieving maximum benefit at threshold 0.795 with 91.11% training accuracy. Using the same dataset, Spearman redundancy elimination (correlation threshold 0.85) retains 49 features. The 135×50 feature matrix (including labels) serves as input for SVM_RFE ranking. Stepwise regression modeling reveals optimal performance at 42 features through 10-fold cross-validation on the training set, with 97.78% internal validation accuracy—thus selecting these 42 features for final modeling. Similarly, SVM_RFE on 330 initial features without Spearman preprocessing achieves optimal performance at 32 features with 97.78% internal validation accuracy, selected for traditional SVM modeling. Detailed results appear in Table 2 .

4.3 Model Evaluation and Comparison

The validation experiment comprises three stages: data preprocessing, feature reduction/modeling, and model evaluation/comparison. After evaluating all three models on validation samples, the improved SVM model incorporating Spearman redundancy elimination demonstrates optimal performance on 59 validation cases, achieving 89.83% classification accuracy.

First, model m is built using Gaussian kernel SVM on 32 features selected

by SVM_RFE. Model m is constructed using the same kernel on 42 features selected by Spearman+SVM_RFE. Model m uses logistic regression on 10 features selected by lasso. To ensure adequate fitting and generalization capabilities, both training and validation data are input to all three models for comprehensive evaluation via ROC curves, accuracy, AUC, sensitivity, and specificity metrics. Figures 6 [Figure 6: see original paper] and 7 [Figure 7: see original paper] show ROC curves for the three models on training and validation sets. Training set predictions show very similar and excellent performance across all models, indicating strong fitting capabilities with minimal differences. Validation set ROC curves all lie far from the diagonal axis, demonstrating good generalization, but with noticeable differences: model m ' s curve is enveloped by m , indicating m ' s superior validation performance. SVM-based models intersect with the logistic regression radiomics model, requiring additional metrics for comparison.

For further generalization comparison, Table 3 details each model' s accuracy, AUC, sensitivity, and specificity. Since SVM' s hyperplane maintains equal margins to different class support vectors without threshold variation, its ROC curve differs from logistic regression with fewer inflection points, making AUC comparison difficult. However, SVM shows significant advantages in prediction accuracy and clinical sensitivity. To further demonstrate m ' s superiority, clinical decision curves visualize net decision benefits across models in Figure 8 [Figure 8: see original paper].

Table 3: Evaluation Metrics of Three Models on Validation Set

Model	Accuracy	AUC	Sensitivity	Specificity
Lasso+LR	84.75%	-	-	-
Spearman+RFE+SVM	89.83%	-	-	-
SVM_RFE+SVM	84.75%	-	-	-

Figure 8 shows that across threshold probability ranges of 0.2–0.85, model m ' s net decision benefit significantly exceeds the other two models, with a continuous and sufficiently large optional threshold probability range—demonstrating excellent judgment criteria. In summary, experiments prove that the SVM model combining Spearman redundancy elimination and SVM_RFE dimensionality reduction effectively supports the core evaluation component of the proposed tumor disease overtreatment and misdiagnosis control mechanism, providing beneficial supervision to reduce the incidence of medical errors during tumor diagnosis and treatment.

References

[1] Chen Wanqing, Zheng Rongshou, Baade P D, et al. Cancer statistics in China, 2015 [J]. CA Cancer J Clin, 2016; 66 (2): 115-32.

[2] Lu Zhifang. On the regretful death of Professor//Dr. Shi Ying-Kang [J]. Quant Imaging Med Surg., 2016, 6 (3): 334-337.

[3] Kallianidou K, Kiakou M, Tsoukalas N, et al. Medication administration in hospital: Difficulties and errors [J]. Arch Hellen Med, 2017, 34 (1).

[4] Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis [J]. European Journal of Cancer, 2012, 48 (4): 441-6.

[5] Carneiro G, Oakdenrayner L, Bradley A P, et al. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography [C]// Proc of the 14th IEEE International Symposium on Biomedical Imaging. 2017.

[6] Liu Wenhua. Study on medical treatment behavior check program based on E-health collaborative diagnosis and treatment platform [D]. Shantou: Shantou University, 2015. 1-42.

[7] Zhou Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016: 121-125.

[8] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46 (1): 389-422.

[9] Huang M L, Hung Y H, Lee W M, et al. SVM-RFE based feature selection and taguchi parameters optimization for multiclass SVM classifier [J]. Scientific World Journal, 2014, 2014: 795624: 1-10.

[10] Liu Yang, Zhang Zhou, Zhou Qinglei. Study on fuzzy rough set rule mining method for medical health data [J]. Computer Science, 2014, 41 (12): 164-167.

[11] Huang Y Q, Liang C H, He L, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer [J]. Science Foundation in China, 2016, 34 (4).

[12] Pal M, Foody G M. Feature selection for classification of hyperspectral data by SVM [J]. IEEE Trans on Geoscience & Remote Sensing, 2010, 48 (5): 2297-2307.

Note: Figure translations are in progress. See original paper for figures.