

# Research on Deep Mining of Medical and Health Data and Disease Prediction Based on Machine Learning

Yuxi Liu

*College of St Hild & St Bede, Durham University, Durham, United Kingdom*

**Abstract:** With the rapid advancement of medical informatization, massive multi-source medical and health data have provided an unprecedented data foundation for disease prediction. This paper systematically analyzes the type classification, fusion processing and feature engineering of medical and health data, constructs a disease-prediction model system centered on ensemble learning, and introduces interpretability methods such as SHAP to realize model transparency. Multi-task prediction frameworks are designed for chronic diseases and acute infectious diseases respectively, and model performance is evaluated using metrics including AUC and F1-score together with multi-fold cross-validation. Experimental results show that the ensemble learning method achieves an AUC above 0.90 in both diabetes and pneumonia prediction tasks, approximately 8% higher than single baseline models. Interpretability outputs effectively assist clinical decision-making. Machine-learning-driven deep mining of medical data possesses remarkable clinical translational value and is expected to advance disease prediction from empirical judgment to precise quantitative decision-making.

**Keywords:** Machine Learning; Medical and Health Data; Disease Prediction; Ensemble Learning; Interpretability

## 1. Introduction

The healthcare sector is undergoing profound digital transformation. Electronic medical records, wearable devices, genomics and public-health surveillance systems generate large-scale structured and unstructured data every day, forming a big-data ecosystem of medical and health data characterized by high dimensionality, heterogeneity and temporality. How to extract knowledge with predictive value from such data has become a cutting-edge topic

at the intersection of medical informatics and artificial intelligence<sup>[1]</sup>. Traditional statistical methods exhibit obvious limitations when handling high-dimensional nonlinear data. In contrast, machine learning offers a new paradigm for early-stage disease prediction owing to its powerful pattern-recognition and adaptive-modeling capabilities<sup>[2]</sup>. Promoting deep mining of medical and health data and the construction of disease-prediction systems carries important practical and academic significance for improving public-health prevention-and-control capacity and achieving precision-medicine goals<sup>[3]</sup>.

## 2. Fundamentals of Deep Mining of Medical and Health Data for Disease Prediction

### 2.1 Classification of Medical and Health Data and Analysis of Deep-Mining Value

Medical and health data can be divided into four major categories according to source and structure: clinical diagnosis-and-treatment data (electronic medical records, laboratory-test and examination reports, imaging data), public-health data (disease-surveillance reports, epidemic direct-reporting systems), individual health-management data (vital signs collected by wearable devices, health archives), and genomic and omics data. Different data types deliver complementary mining value: clinical data directly reflect disease-progression patterns; public-health data reveal population epidemic trends; vital-sign data support early-warning functions; and omics data provide insights into disease-causing mechanisms at the molecular level. It should be emphasized that prediction performance using a single-data source is markedly limited. Collaborative mining of multimodal data can address the information blind spots in individual modalities to build a complete prediction chain for individual risk assessment and population trends. The strength of deep mining is also its ability to uncover hidden factors and non-linear relationships in the

data that traditional medical statistics fail to detect, providing new data-driven directions for the study of disease pathology.

Of particular note, with the establishment of large-scale longitudinal cohort databases (such as the UK Biobank, which contains genetic and health records of more than 500,000 people), deep mining of medical data has evolved from a one-off static cross-sectional analysis to dynamic temporal modelling. Continuous tracking of an individual's health changes can detect early warning signs of illness years in advance and expand the time window for intervention. Under this new paradigm of "process-based early warning" instead of "outcome diagnosis", machine learning for medical data mining can now be applied.

## 2.2 Fusion Processing and Quality Optimisation Methods for Multi-source Medical Data

The primary problems of multi-source medical data fusion include lack of uniform standards, substantial missing data and noise interference [4]. At the data preprocessing stage, based on the mechanism of missing data (MCAR/MAR/MNAR), choose an imputation strategy; use multiple imputation for randomly missing values and reasonable estimation that incorporates medical prior knowledge for systematically missing values.

For standardization of heterogeneous data, cross-institutional data alignment is realized by establishing unified data-element dictionaries and coding specifications (e.g. ICD-10 diagnostic codes, LOINC laboratory-test codes). To address class imbalance (scarce disease-positive samples), SMOTE oversampling is combined with cost-sensitive learning for correction. Data-quality optimization also incorporates outlier-detection mechanisms, whereby unsupervised methods such as Isolation Forest identify noisy clinical records. Datasets systematically processed via fusion show substantial improvements in feature completeness and sample representativeness, laying a solid foundation for subsequent modeling.

## 2.3 Feature-Extraction and Variable-Selection Strategies for Machine-Learning Modeling

Feature engineering serves as the critical link between raw medical data and machine-learning models. In feature extraction, statistical descriptors such as trend slope and fluctuation

variance are extracted from time-series vital-sign data; TF-IDF or medical named-entity recognition is adopted for textual medical-record data to extract structured semantic features; convolutional neural networks extract high-level visual features from imaging data. A three-stage "filter-wrapper-embedding" joint framework is used for variable selection: redundancy filtering of variables is performed based on mutual information and variance thresholds at the filter stage; recursive feature elimination (RFE) finds an optimal feature subset in the wrapper stage; and sparse screening is conducted by Lasso regularization and built-in random forest feature importance in the embedding stage. It is an easy-to-use and reasonably precise screening function. Reduce the high-dimensional space of medical features to a small number of clinically interpretable variables to prevent overfitting and maintain predictive power.

Additionally, for the ubiquitous temporal features in electronic health records (e.g., repeated outpatient visits and continuously monitored metrics), sliding-window statistics and difference transformations can be employed to convert dynamic trends into static feature vectors, thus capturing information about the speed of disease progression.

For textual chief complaints and discharge summaries, medical semantic encoding using BERT-family pre-trained language models can markedly improve the utilization of unstructured information and remedy the inherent limitation of structured features, which inadequately describe clinical details.

## 3. Machine-Learning Method System for Deep Mining of Medical and Health Data

### 3.1 Principles and Application Scenarios of Typical Machine-Learning Algorithms for Medical-Data Mining

Common machine-learning algorithms for medical-data mining include logistic regression, support-vector machines, decision trees, random forests and gradient-boosting trees. Thanks to its linear interpretability, logistic regression applies to risk-score modeling for low-dimensional structured clinical data. By means of kernel-function mapping, support-vector machines (SVM) handle nonlinear classification problems and deliver stable performance for small-sample high-dimensional genomic

datasets. Decision trees feature intuitive interpretability, enabling clinicians to understand decision paths, yet they are sensitive to data noise and prone to overfitting. Random forests construct multiple decision trees under the Bagging strategy and integrate prediction outputs, yielding enhanced generalization capacity especially for multi-dimensional electronic-medical-record datasets containing redundant features<sup>[5]</sup>. Gradient-boosting trees (e.g. XGBoost, LightGBM) iteratively correct residual errors under the Boosting mechanism and generally achieve state-of-the-art performance for classification and regression tasks on structured medical data. Key differences in applicable scenarios among algorithms lie in data scale, feature dimensionality, sample balance and clinical interpretability requirements. The ensemble-prediction output of a random-forest model is defined as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T h_t(x)$$

where  $T$  denotes the number of base learners (decision trees),  $h_t(x)$  represents the prediction output of the  $t$ -th tree for an input sample, and  $\hat{y}$  is the final ensemble prediction result. Multi-tree voting is employed to reduce variance and improve the generalisation stability of medical data.

Tabular data are suitable for the neural network methods that have gained popularity recently, such as TabNet. Attention mechanisms are used to dynamically select features, and the performance on some large-scale medical datasets is comparable to that of gradient-boosted trees. However, these models need a large amount of training data, are very sensitive to hyperparameter tuning, and are less stable than ensemble tree models in small-sample clinical environments. Therefore, in the selection of an algorithm, only one performance indicator should not be used; all of the above factors need to be considered at the same time.

### 3.2 Construction and Performance Optimization of Disease Prediction Models Based on Ensemble Learning

Ensemble learning combines several base learners to reduce the bias-variance trade-off and has achieved good results in disease prediction. Model construction is in three stages. First, the quality-improved multi-source medical data are divided into a training set, a validation set and a test set at a ratio of 7:1.5:1.5.

Second, a Boosting ensemble framework is built with XGBoost as the backbone; random forests and logistic regression are simultaneously

introduced as diverse base learners. Finally, under the Stacking strategy, predicted probabilities output by the validation set are taken as meta-features to train a secondary meta-learner for final fusion. To enhance performance, Optuna is used to perform Bayesian optimisation of hyperparameters, such as tree depth, learning rate and feature sampling rate; the speed improvement over grid search is approximately 40 per cent. In addition, an early-stopping mechanism and L2-regularisation are used to prevent overfitting. To address class imbalance in the loss function and improve the model's recognition of disease-positive samples, sample weight correction has been introduced.

### 3.3 Introduction and Applications of Model Interpretability Methods in Medical and Health Data Mining

The application of machine learning in medicine for diagnosis and prognosis needs to be interpretable; otherwise, doctors will not use it. SHAP (SHapley Additive exPlanations) assigns the marginal contributions of each feature to the prediction as additive effects based on Shapley values from cooperative game theory, and provides dual-level global and local explanations. SHAP-based feature importance ranking has also identified important medical factors in the global prediction of disease (e.g., blood glucose level, age, BMI). Locally, Force Plots show the risk-driving factors of individual patients intuitively. SHAP is used to obtain explanations for all data points, and LIME is used to provide simple local explanations of individual predictions. The first component of SHAP values is the Shapley value: where  $F$  is the whole set of features,  $F_i$  is the subset without feature  $i$ , and  $S$  is the set used for model prediction. This formula shows the average marginal contribution of feature  $i$  over all feasible feature subsets and meets the criteria of efficiency, symmetry, and the dummy variable condition; see Figure 1.

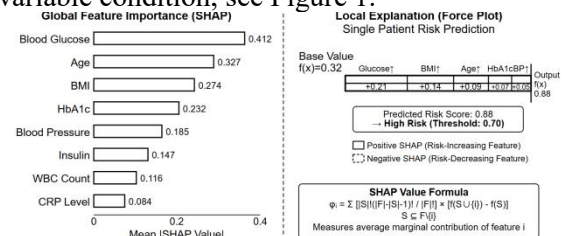


**Figure 1. Schematic Diagram of SHAP Feature-Importance Analysis and Local Interpretation**

## 4. Research and Performance Evaluation of

## Disease Prediction via Machine-Learning-Enabled Deep Mining

### 4.1 Construction of Multi-Task Disease-Prediction Models for Chronic and Acute Diseases

Chronic diseases (represented by type 2 diabetes) and acute diseases (represented by pneumonia) differ substantially in data characteristics, onset temporality and prediction objectives. Therefore, a Multi-Task Learning (MTL) framework is employed for joint modelling. The MTL framework has a shared bottom-layer feature extraction module to obtain common baseline health characteristics of both disease groups (age, BMI, vital signs, basic lab values). Task-specific independent layers are established for each disease to extract different predictive signals: chronic disease modeling focuses on long-term trend variables (rate of change in HbA1c, growth trajectories of waist circumference), and acute disease modeling concentrates on short-term abnormal signals (sudden temperature rise, C-reactive protein peak, sharp increase in leukocyte count). Joint multi-task training achieves knowledge sharing via inter-task regularisation constraints to reduce overfitting caused by a lack of single-task data and significantly improves the prediction stability of small-sample acute disease tasks. The dataset combines outpatient and inpatient data from several Grade-A tertiary hospitals. After standardisation, the total number of cases in the sample exceeds 120,000; therefore, a suitable model can be built.

### 4.2 Evaluation-Metric System and Generalization-Capability Validation for Disease-Prediction Models

A rigorous evaluation system is essential for verifying the clinical practical value of models. The evaluation-metric system takes AUC-ROC (Area Under the Receiver-Operating-Characteristic Curve) as the primary metric, supplemented by F1-score (balancing precision and recall), Sensitivity, Specificity and Brier Score (probability-calibration quality) to form a multi-dimensional evaluation matrix. Sensitivity primarily measures the risk of missed disease diagnoses, while Specificity controls costs of over-diagnosis; trade-offs between the two should be adjusted according to disease severity. Generalization-capability validation employs

dual-validation mechanisms: stratified 5-Fold Cross-Validation (Stratified 5-Fold CV) and external independent test-set validation, avoiding artificially inflated performance stemming from data-distribution discrepancies. External-validation-set data are sourced from hospitals in different geographic regions to test cross-center generalization capacity; see Table 1.

**Table 1. Performance Comparison of Major Models on Diabetes-Prediction and Pneumonia-Prediction Tasks**

| Model               | Task                 | AUC   | F1-score | Sensitivity | Specificity |
|---------------------|----------------------|-------|----------|-------------|-------------|
| Logistic Regression | Diabetes Prediction  | 0.812 | 0.769    | 0.793       | 0.825       |
| Random Forest       | Diabetes Prediction  | 0.873 | 0.831    | 0.845       | 0.896       |
| XGBoost             | Diabetes Prediction  | 0.912 | 0.878    | 0.869       | 0.921       |
| MTL Ensemble        | Diabetes Prediction  | 0.931 | 0.902    | 0.887       | 0.938       |
| Logistic Regression | Pneumonia Prediction | 0.798 | 0.741    | 0.812       | 0.768       |
| Random Forest       | Pneumonia Prediction | 0.856 | 0.819    | 0.871       | 0.834       |
| XGBoost             | Pneumonia Prediction | 0.897 | 0.861    | 0.903       | 0.876       |
| MTL Ensemble        | Pneumonia Prediction | 0.918 | 0.886    | 0.921       | 0.901       |

### 4.3 Clinical Translation and Decision-Support Applications of Machine-Learning-Based Disease-Prediction Outcomes

Clinical translation of model predictions must bridge the gap between technical accuracy and clinical usability. For decision support, model-output disease-risk probabilities together with SHAP feature-contribution values are presented on clinical-assistance-decision interfaces. Clinicians obtain risk scores while intuitively identifying concrete clinical variables driving such risks, thus achieving "interpretable early-warning alerts" rather than "black-box alarms". In risk-stratification applications, patients are divided into low-, medium- and high-risk groups according to predicted probabilities. The high-risk group will be automatically referred for further examination and consultation with a doctor; the medium-risk group will enter dynamic follow-up management; and the low-risk group will receive regular health management. A dynamic update mechanism can incrementally update the online model with new clinical data and continue

iterative optimisation according to changes in clinical practice; see Figure 2.

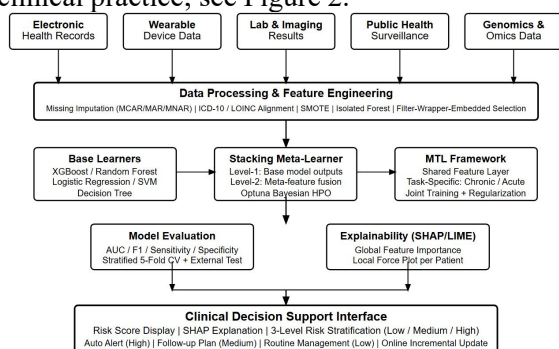


**Figure 2. Flowchart of Machine-Learning-Driven Disease-Prediction Clinical Decision-Support System**

## 5. Conclusion

With the continuous improvement of medical informatics infrastructure, machine learning and big data have been deeply integrated in recent years to advance the development of personalised medicine and preventative health care. Improve the quality of multi-source data fusion, build multi-task ensemble learning frameworks and introduce interpretability technology to construct disease prediction systems with both strong predictive power and clinical reliability. The current model performance meets the basic requirements for auxiliary clinical applications; however, embedding algorithms into diagnosis-and-treatment workflows and constructing a prediction-feedback-iteration closed loop are still bottlenecks to real-world deployment. In the

future, federated learning will be used to address cross-institutional data privacy problems, improve predictions for rare diseases and infectious disease outbreaks, advance medical machine learning from "usable" to "practical", and shift from auxiliary diagnosis to proactive prevention.

## References

- [1] Lu J X, Ren C L, Sun Q. Research on machine-learning optimization applications in big medical-and-health data[J]. Computer Knowledge and Technology, 2026, 22(14): 64-68.
- [2] Shi X Q, Zhao J, Wang L L, et al. Review on precision-medicine applications based on big-data-analytics technology[J]. Chinese Hospital Management, 2021, 41(5): 26-31.
- [3] Li T H. Construction of an early-prevention-and-control platform for precise prediction of infectious diseases based on health-medical data from urban medical-alliance systems[J]. China Informatization, 2021, (3): 103-105+102.
- [4] Cheng W Q, Gao Z S, Liu Y. Research on collaboration and applications of medical-and-health data based on trusted-data spaces[J/OL]. China Digital Medicine:1-8[2026-09-22]. <https://link.cnki.net/urlid/11.5550.R.20260911.2214.006>.
- [5] Zhang Y Y, Zou X K, Xian L L. New disease-prevention-and-control modes driven by big health-medical data[J]. China Health Industry, 2024, 21(3): 132-135.