NLP-Driven Text Mining of Medical Examination Data: Exploring Potential Associations and Patterns in Dual-High Diseases

Tianyuan Liang
Department of Computer Science
and Engineering
liang.1511@buckeyemail.osu.edu

Johnson Zhong
Department of Computer Science
and Engineering
zhong.820@buckeyemail.osu.edu

Abstract

With the rise of chronic non-communicable diseases due to aging populations and lifestyle changes, hypertension and hyperlipidemia ("dual-high" diseases) have become increasingly common and pose significant health risks. This project explores the use of machine learning, particularly the LightGBM model, to predict the risk of dual-high diseases based on real-world, de-identified physical examination data. We performed extensive data cleaning, structured text processing, and feature engineering, including Doc2Vec-based extraction for long medical texts. The model identifies key risk factors and offers insights into the association between physiological indicators and dual-high diseases, supporting early intervention and personalized prevention.

1 Introduction

With the improvement of living standards and changes in lifestyle, dual-high diseases such as hypertension and hyperlipidemia are becoming increasingly prevalent among the population, posing serious threats to health. NLP-based methods can process vast amounts of physical examination data and uncover potential associations and patterns. This project aims to utilize the NLP knowledge learned in class to perform text mining on physical examination data and improve the prediction, prevention, and treatment of related diseases.

2 Problem Statement

Chronic diseases such as hypertension and hyperlipidemia (collectively known as dual-high diseases) pose significant health risks, including cardiovascular disease and stroke. Traditional risk assessment models rely on structured numerical data, often overlooking valuable insights embedded in unstructured medical text such as doctors' notes and diagnostic reports. This project integrates Natural Language Processing (NLP) and Data Mining techniques, leveraging medical text analysis alongside structured physiological data, the system aims to extract meaningful features from unstructured reports, enhance predictive accuracy, and uncover hidden risk factors.

3 Materials and Methods

3.1 Datasets

The data is sourced from the Alibaba Cloud platform and authorized by Meinian Health. It is extracted from its big health biological sample database, including 100,000 de-identified physical examination records and 6,000 genetic data samples. Each record contains annotations for hypertension and hyperlipidemia, along with hundreds of physical examination indicators and genetic loci. All data has been de-identified in strict accordance with internationally recognized medical information anonymization standards. The dataset used in this project consists of real-world, de-identified hospital examination records, including: Structured numerical data (e.g., blood pressure, cholesterol levels). Unstructured textual data (e.g., medical examination reports, diagnostic summaries, medical history). In the data preprocessing work, we need to perform structuring, clean labeled datasets, etc., to ensure data quality and consistency. A rule-based classifier will be used to perform certain domain knowledge-based filtering

3.2 Feature Engineering

We need to perform feature extraction on various types of features, including numerical features, short-text features, and long-text features. The proposed approach involves using label encoding for short-text feature extraction, while Word2Vec is utilized for long-text feature extraction.

3.3 Model Selection and Training

In this project, we selected LightGBM (Light Gradient Boosting Machine) as the primary model for predicting the risk of dual high diseases (hypertension and hyperlipidemia). This decision was made based on LightGBM's well-recognized efficiency, scalability, and high accuracy in handling structured data, especially in scenarios involving a large number of features and sparse representations, such as those encountered in medical datasets.

LightGBM provides faster training speed and lower memory usage by utilizing histogram-based algorithms and optimized leaf-wise tree growth. It supports GPU acceleration, which significantly reduces model training time. Compared with other boosting algorithms like DART and GOSS, GBDT (Gradient Boosted Decision Tree) in LightGBM showed the lowest MSE and highest R² in our task, especially on key indicators such as systolic and diastolic blood pressure and serum lipid profiles.

The model was trained using a 5-fold cross-validation strategy to ensure generalization. Input features included a combination of cleaned numerical features and vectorized text features (from long medical notes), where long text was encoded using Doc2Vec. During training, the top 250 features (about 37% of all features) were selected based on prior importance ranking.

3.4 Model Evaluation and Interpretation

Metrics: MSE (Mean Squared Error), R² (R-squared), AUC (Area Under Curve). Explainability Analysis:Feature importance ranking to identify key risk factors.

4 Expected Outcome

A machine learning model that improves disease risk prediction by integrating NLP-based insights from medical reports.

A structured approach for mining medical data, combining text analysis and numerical feature extraction. Improved explainability of disease risk factors, supporting early intervention and personalized treatment recommendations.

5 Related Work

With the rapid advancement of information technology, artificial intelligence (AI) has been increasingly applied to chronic disease prediction and management. Numerous studies have demonstrated the effectiveness of machine learning (ML) algorithms, especially ensemble methods, in predicting cardiovascular diseases.

Yang et al.(Yang et al., 2023)proposed an enhanced artificial neural network (ANN) using the Framingham Heart Study dataset for coronary heart disease (CHD) prediction. Their model surpassed the Framingham Risk Score (FRS) in sensitivity and specificity, though it had a lower AUC. Du et al.(Qin, 2023) applied machine learning to predict CHD in hypertensive patients based on electronic health records, with XGBoost achieving the highest AUC of 0.943.

Akella and Akella (Dhruva Kumar et al., 2022)applied six ML algorithms on the Cleveland dataset for coronary artery disease (CAD) prediction, with neural networks achieving over 93% accuracy. Muhammad et al. [18] used clinical data from two Nigerian hospitals and found random forests yielded the best performance with 92.04% accuracy.

Wang et al.(Wang et al., 2014)combined logistic regression and ANN for hypertension prediction using BRFSS data, achieving over 72% accuracy and an AUC above 0.77. You et al.(You et al., 2023) utilized SpO2 signals to predict hypertension risk in obstructive sleep apnea (OSA) patients. Their random forest model reached 84.4% accuracy and outperformed clinical expert predictions.

Qin(Qin, 2023) developed an XGBoost-based model using data from L Hospital, achieving a 12.7% average error rate, further validating XGBoost's effectiveness in hypertension prediction. Collectively, these studies highlight the growing role of machine learning—particularly boosting algorithms like XGBoost and LightGBM—in enhancing predictive accuracy for chronic conditions such as cardiovascular disease and hypertension. Building upon this foundation, the present study explores the use of LightGBM to predict dual-high diseases (hypertension and hyperlipidemia), aiming to support early detection in at-risk populations.

Most existing research relies on a single data source and has a small sample size. In contrast, our study includes approximately 60,000 records, with nearly 300 features retained after feature selection, making the data more comprehensive. We use multiple models for disease prediction, primarily employing LightGBM. Unlike most prediction tasks, our research innovates in multi-target space preprocessing based on electronic medical record (EMR) data, aiming to standardize the metrics across different data sources. Text features are vectorized using the Doc2Vec model, and predictions are ultimately made using the LightGBM model. We also use the Mean Decrease Impurity (MDI) method for global importance assessment to enhance the model's interpretability. The highly correlated features are then input into the LLM for inference related to their association with hypertension and hyperlipidemia diseases, supplemented by literature or medical record databases (RAG), resulting in an analysis of potential disease risk factors.

We also plan to combine the EMR dataset and the obtained model with large language models (LLMs) to generate inference-based risk predictions for doctors. Doctors can combine their personal expertise with the generated medical history information to provide more interpretable decision support for patients. Our research approach is to enable the LLM to learn the highly correlated features and corresponding EMR data derived through the MDI method, and extract structured text information similar to what we input into the Doc2Vec model from doctor-patient dialogues. During this process, we supplement the inference results with RAG, then vectorize the results, use the trained model for indicator prediction, and finally feed the predicted results back into the LLM to generate inference-based risk predictions for doctors' reference. The ultimate goal of this approach is to reduce the doctors' workload and guide patients toward healthier living.

6 Results of Conducted Experiments

6.1 Transforming Raw Data into Structured Data

In terms of the dataset, there are 8,104,368 unstructured data entries, each representing a physical examination result. The examination project ID field has identical values for the same examination item. The examination result field includes both numerical and character data types, and the results are provided in an unstructured text format. The meaning of each data entry is that the patient (vid) obtained the examination result (field_results) for the examination item (table_id).

A dataset of 38,198 labeled entries, with labels consisting of five indicators representing blood pressure and blood lipid levels: systolic blood pressure, diastolic blood pressure, serum triglycerides, serum high-density lipoprotein (HDL), and serum low-density lipoprotein (LDL).

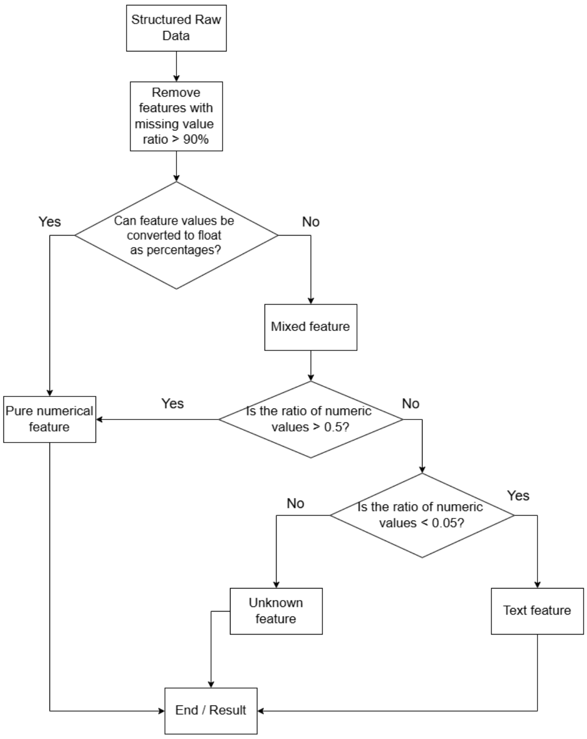

Structured Raw Data
Remove features with missing value ratio > 90%
Yes
Can feature values be converted to float as percentages?
No
Mixed feature
Pure numerical feature
Yes
Is the ratio of numeric values > 0.5?
No
No
Is the ratio of numeric values < 0.05?
Yes
Unknown feature
Text feature
End / Result

Figure 1: Feature Processing Flowchart

6.2 Structured Raw Data Processing

For features classified as numerical but still containing a small proportion of text feature values, we calculate the proportion of text feature values. If the proportion of text feature values is less than 5%, we directly drop these feature values. If the proportion is greater than 5%, we manually convert them into numerical values. For text features, we performed operations such as cleaning punctuation, data standardization, and replacing mappings with medical knowledge. We also used the jieba segmentation tool for word segmentation and introduced a medical dictionary to assist with the segmentation. For features with a length of 6 or fewer characters, we considered them short-text features; otherwise, they are considered long-text features. Among the 145 text features, 100 are long-text features, and 45 are short-text features.The flow of feature processing is shown in figure 1.

For the extraction of short-text features, inspired by one-hot encoding, we adopted a method similar to label encoding. The basic idea is to map each category to a unique integer label. For long-text features, initially, we used the TF-IDF extraction method, but the results were unsatisfactory. The extracted feature file was as large as 4GB (the original file was only 100MB), and it contained a very high proportion of zero values. After reconsideration,

Feature	Doc2Vec	Word2Vec
Input Data	Whole document	Single words
Context Info	Uses document-level context	Uses surrounding words
Representation	Vector of document	Vector of each word
Model Type	PV: DM + DBOW	CBOW and Skip-gram
Training Target	Match doc and context words	Match target with context words
Document Length	Handles varying lengths	Fixed-length may be needed
Similarity Comparison	Compares full documents	Compares individual words

Table 1: Simplified comparison between Doc2Vec and Word2Vec models

Hyperparameters	Value
vector_size	5
window	5
min_count	1
sample	0
workers	16
hs	0
dm	1
negative	5
dbow_words	1
dm_concat	1
epochs	10

Table 2: Hyperparameters for Doc2Vec.

we switched to the Doc2Vec method and set the vector size (vectorsize) to 5, meaning we believe that five components can reasonably represent the patient's examination results on these five indicators. This method successfully maps the complex examination data into a high-dimensional yet concise vector space, making it easier for analysis and comparison.

Doc2Vec is an extension of Word2Vec. Word2Vec is a method for representing words as dense vectors that can capture semantic relationships between words. Doc2Vec builds upon this by training a neural network model to learn vector representations of entire documents, enabling the comparison of document similarities within a continuous vector space. Its ability to extract features from long texts is achieved through the Paragraph Vector model, a technique designed for learning document-level representations. The Paragraph Vector model extends Word2Vec to handle entire documents rather than just individual words.(The comparison between Doc2Vec and Word2Vec is shown in table 1, the hyperparameters for Doc2Vec is shown in table 2).

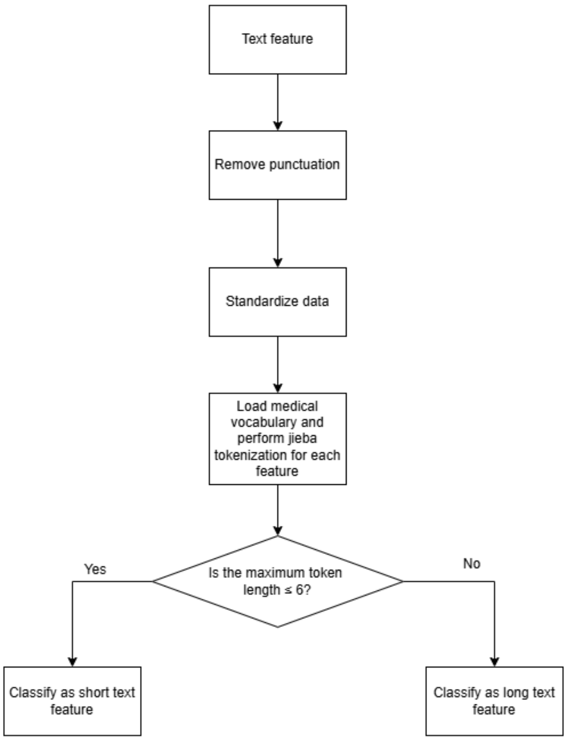

Text feature
Remove punctuation
Standardize data
Load medical vocabulary and perform jieba tokenization for each feature
Is the maximum token length ≤ 6?
Yes
No
Classify as short text feature
Classify as long text feature

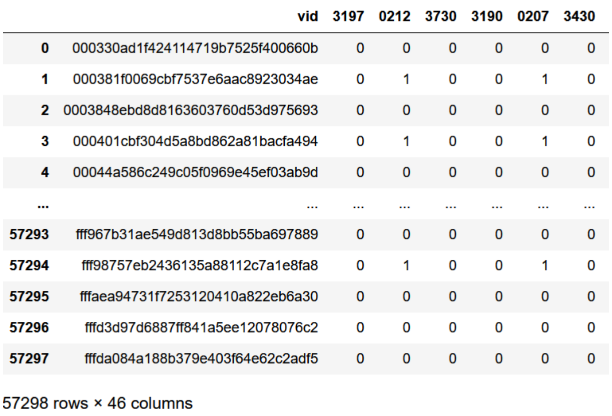
Figure 2: Text Feature Extraction Process

6.3 Prediction Model Based on LightGBM

We used KNN and LightGBM for model training. The results obtained from KNN were not very satisfactory, while the model obtained from LightGBM was more promising.

This study selects LightGBM as the prediction model. Developed by Microsoft, LightGBM is a boosting ensemble model that serves as an optimized and efficient implementation of the Gradient Boosting Decision Tree (GBDT) algorithm. Although it shares conceptual similarities with models like XGBoost, LightGBM demonstrates superior performance in various aspects.

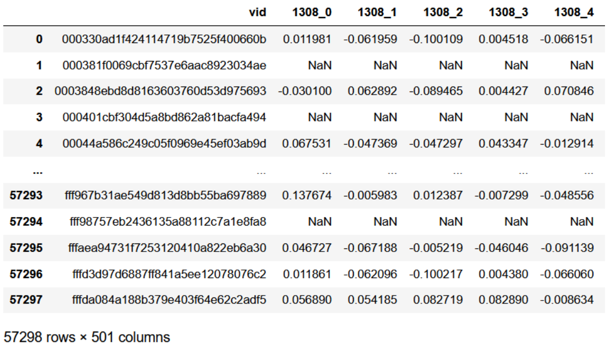
According to its official documentation, LightGBM offers several advantages, including faster training speed, lower memory usage, higher accu-



	vid	3197	0212	3730	3190	0207	3430
0	000330ad1f424114719b7525f400660b	0	0	0	0	0	0
1	000381f0069cbf7537e6aac8923034ae	0	1	0	0	1	0
2	0003848ebd8d8163603760d53d975693	0	0	0	0	0	0
3	000401cbf304d5a8bd862a81bacfa494	0	1	0	0	1	0
4	00044a586c249c05f0969e45ef03ab9d	0	0	0	0	0	0
...	...	...	...	...	...	...	...
57293	fff967b31ae549d813d8bb55ba697889	0	0	0	0	0	0
57294	fff98757eb2436135a88112c7a1e8fa8	0	1	0	0	1	0
57295	fffaea94731f7253120410a822eb6a30	0	0	0	0	0	0
57296	fffd3d97d6887ff841a5ee12078076c2	0	0	0	0	0	0
57297	fffda084a188b379e403f64e62c2adf5	0	0	0	0	0	0

57298 rows × 46 columns

Figure 3: Short text features after label encoding



	vid	1308_0	1308_1	1308_2	1308_3	1308_4
0	000330ad1f424114719b7525f400660b	0.011981	-0.061959	-0.100109	0.004518	-0.066151
1	000381f0069cbf7537e6aac8923034ae	NaN	NaN	NaN	NaN	NaN
2	0003848ebd8d8163603760d53d975693	-0.030100	0.062892	-0.089465	0.004427	0.070846
3	000401cbf304d5a8bd862a81bacfa494	NaN	NaN	NaN	NaN	NaN
4	00044a586c249c05f0969e45ef03ab9d	0.067531	-0.047369	-0.047297	0.043347	-0.012914
...	...	...	...	...	...	...
57293	fff967b31ae549d813d8bb55ba697889	0.137674	-0.005983	0.012387	-0.007299	-0.048556
57294	fff98757eb2436135a88112c7a1e8fa8	NaN	NaN	NaN	NaN	NaN
57295	fffaea94731f7253120410a822eb6a30	0.046727	-0.067188	-0.005219	-0.046046	-0.091139
57296	fffd3d97d6887ff841a5ee12078076c2	0.011861	-0.062096	-0.100217	0.004380	-0.066060
57297	fffda084a188b379e403f64e62c2adf5	0.056890	0.054185	0.082719	0.082890	-0.008634

57298 rows × 501 columns

Figure 4: Document vectorization using Doc2Vec

racy, support for parallel learning, the ability to handle large-scale datasets, and native support for categorical features.

These advantages are mainly attributed to the following technical implementations. First, LightGBM uses a histogram-based algorithm during training to discretize continuous floating-point features into a fixed number of bins and builds histograms to efficiently determine the best split points. This significantly improves training speed and reduces memory consumption. Second, it adopts the Gradient-based One-Side Sampling (GOSS) method, which selects data instances with large gradients and randomly samples from the rest, thereby reducing computational cost without sacrificing accuracy. Third, LightGBM applies a leaf-wise tree growth strategy with depth limitation. Instead of growing the tree level-by-level, it selects the leaf node with the highest gain for splitting at each step, which results in better accuracy for the same number of splits. Finally, multi-threading optimization is utilized to further accelerate training.

Unlike traditional GBDT tools that adopt level-wise tree growth, LightGBM uses a leaf-wise growth strategy with a maximum depth constraint. In each iteration, it identifies and splits the leaf node with the highest gain, which is usually the node with the largest number of data samples. This approach significantly reduces loss and improves prediction accuracy. However, it can also lead to very deep trees and potential overfitting. To address this, LightGBM introduces a tree depth limit to control model complexity and enhance generalization.

During model training, a relatively large learning rate such as 0.05 or 0.04 led to excessively fast convergence, making it difficult for the model to achieve optimal results. However, when the learning rate was reduced to 0.02 or 0.015, the model performance did not improve, and the training time became significantly longer. Specifically, when the learning rate was set to 0.015, the total training time for five models reached approximately 35 hours. As a result, a learning rate of 0.025 was ultimately adopted.

The best results were obtained by using gbdt as the boosting type for gradient boosting trees. The loss function and evaluation metric were both set to mean squared error (MSE). Considering that the number of features remained relatively high even after data cleaning—and that some features contributed very little to the model (e.g., prostate, uterus)—the feature vector dimensionality reached nearly 700 after vectorization. To prevent overfitting, 60% of the features were randomly selected for training in each iteration. Additionally, a relatively large L1 regularization coefficient of 0.3 was applied.

Regarding the choice of `boosting_type`, LightGBM supports several options, each with distinct mechanisms and advantages:

(1) gbdt (Gradient Boosting Decision Tree) is the default and most commonly used method. It follows the traditional gradient boosting framework, where each new tree is trained to fit the residual errors of the ensemble model so far. Mathematically, the residuals can be expressed as the negative gradients of the loss function L, turning gradient boosting into a functional gradient descent algorithm. At each iteration, LightGBM fits a regression tree to these negative gradients and updates the ensemble accordingly.

(2) dart (Dropouts meet Multiple Additive Regression Trees) is an enhanced version of GBDT that introduces a dropout mechanism inspired by neural networks. During training, some decision

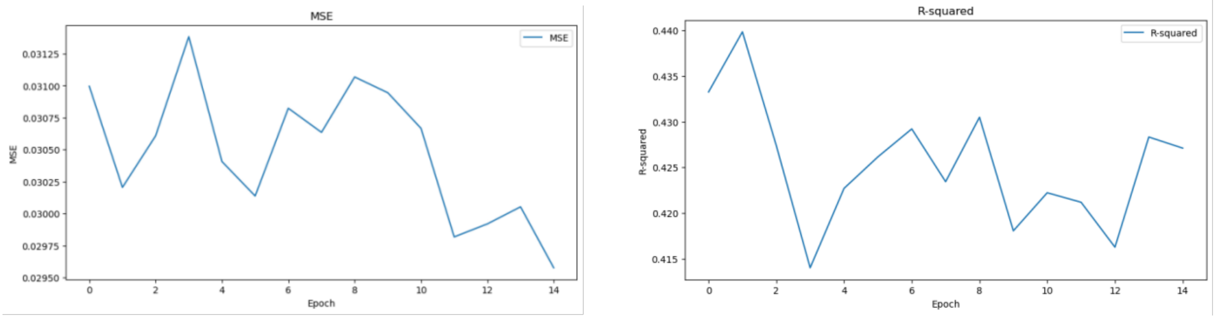


Figure 5: Evaluation metrics for the GBDT model: Left – Mean Squared Error (MSE), Right – R^2 score.

trees in the current ensemble are randomly dropped before fitting the next tree. This stochastic regularization technique helps prevent overfitting by reducing the reliance on specific trees and encouraging the model to be more robust.

(3) goss (Gradient-based One-Side Sampling) is a unique acceleration technique designed specifically for LightGBM. It ranks training instances by the absolute values of their gradients and retains those with larger gradients (which usually carry more information about the loss), while randomly discarding a portion of the instances with smaller gradients. This significantly reduces the dataset size used in each iteration, thereby speeding up training while maintaining model accuracy.

Each of these boosting types balances training speed, regularization strength, and predictive performance differently. In practice, gbdt offers strong general-purpose performance, while dart is more suitable in scenarios prone to overfitting, and goss is ideal when training speed is critical on large datasets.

To evaluate the impact of different boosting strategies on model performance, three variants of LightGBM's boosting_type were tested: gbdt, dart, and goss. For each configuration, two evaluation metrics were selected: Mean Squared Error (MSE) and the coefficient of determination (R^2 score).

Figure 5 illustrates the evaluation results for the gbdt model, with the left plot depicting MSE and the right plot showing R^2. These metrics provide insight into both the accuracy of the predictions and the model's ability to explain the variance in the data.

Subsequently, the performance of the dart model(Figure 6) is visualized using the same metrics. While dart incorporates dropout-based regularization to reduce overfitting, it did not outperform the baseline gbdt model in this specific task.

Boosting Type	**Parameters**	**MSE**	**R^2**
gbdt	LDL	0.0305	0.3878
dart	LDL	0.6490	-12.0505
goss	LDL	0.0391	0.2151

Table 3: Hyperparameters for Doc2Vec.

The evaluation plots for goss are also presented(Figure 7). Although goss achieved better performance than dart by focusing on high-gradient samples, its accuracy still lagged behind that of gbdt.

Based on this comparative analysis(Table 4), gbdt was selected as the most suitable boosting_type for the final model due to its superior performance across all evaluation metrics. The resulting performance is summarized as follows:

After training with the top 30% of features and tuning the parameters, we obtained the the following ideal parameter list in Table 6. After training for 15 epochs, the evaluation metrics for each index model are summarized in Table 5.

To further investigate the individual contributions of textual and numerical features, we conducted additional experiments by training two separate LightGBM models: one using only structured numerical features, and another using only unstructured textual features (processed via Doc2Vec). The performance of these single-modality models was then compared to the previously discussed combined model that utilizes both feature types.

Across the five key prediction targets—Systolic Blood Pressure, Diastolic Blood Pressure, Serum Triglycerides, Serum HDL, and Serum LDL—we observed that the combined model consistently outperformed both text-only and number-only models in terms of Mean Squared Error (MSE). This trend highlights the synergistic benefit of integrating both structured and unstructured data sources. Notably, the MSEs from the text-only model were 0.01 to

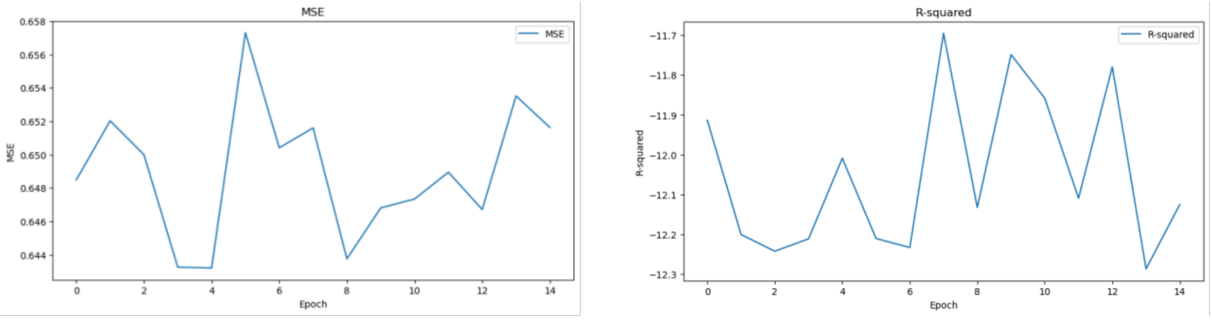


Figure 6: Evaluation metrics for the DART model: Left – Mean Squared Error (MSE), Right – R^2 score.

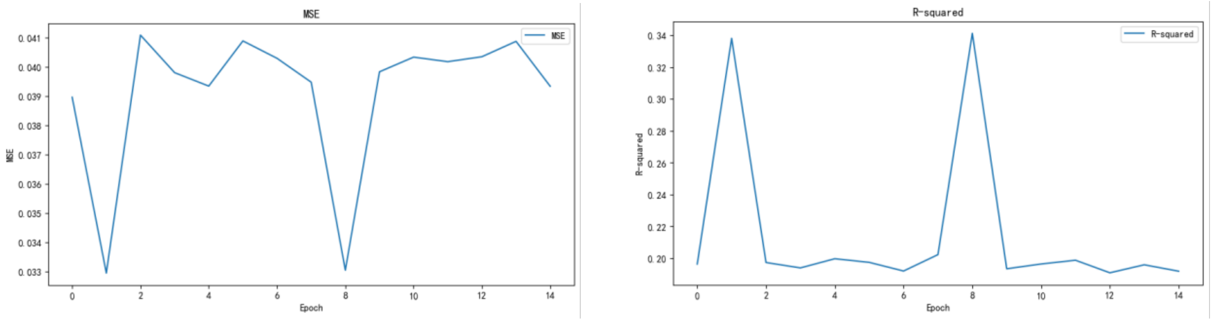


Figure 7: Evaluation metrics for the GOSS model: Left – Mean Squared Error (MSE), Right – R^2 score.

0.02 higher than the combined model, suggesting that while unstructured textual information derived from medical notes contains valuable context, it lacks the quantitative precision necessary for highly accurate predictions when used in isolation.

On the other hand, the number-only model showed only a slight performance drop, with MSE values approximately 0.001 higher than those of the combined model. This narrow gap underscores the critical role of numerical features—such as blood pressure readings and lipid concentrations—which are directly measurable and have well-established clinical relevance. However, despite their primary predictive power, these numerical features do not fully capture the holistic patient context that textual records can offer.

The improved performance of the combined model, although modest in absolute MSE difference, suggests that textual data contributes subtle, complementary information that helps the model better generalize and refine its predictions. For example, long-form medical notes may include descriptions of lifestyle factors, historical conditions, or physician assessments that are not reflected in raw numerical inputs but still influence patient outcomes. These results reaffirm the value of multi-modal feature integration in medical data modeling tasks, especially in scenarios involving complex, multifactorial health indicators such as dual-high diseases. This integration not only boosts predictive performance but also moves the model closer to the kind of reasoning human clinicians use, incorporating both numbers and narratives for more comprehensive decision-making.

Prediction Parameters	**number**	**text**
Systolic Blood Pressure	0.01530	0.01516
Diastolic Blood Pressure	0.01861	0.01917
Serum Triglycerides	0.07286	0.09557
Serum HDL	0.01132	0.01336
Serum LDL	0.03160	0.04018

Table 4: MSE Comparison of Text-Only and Number-Only Models Across Five Health Indicators.

Prediction Parameters	**MSE**	**R^2**
Systolic Blood Pressure	0.013667	0.373500
Diastolic Blood Pressure	0.017644	0.308722
Serum Triglycerides	0.071266	0.435021
Serum HDL	0.010652	0.425312
Serum LDL	0.030483	0.387840

Table 5: Model prediction results with corresponding average MSE and R^2 values.

Hyperparameters	Value
`learning_rate`	0.025
`boosting_type`	gbdt
`objective`	regression
`metric`	mse
`num_leaves`	60
`feature_fraction`	0.6
`min_data`	100
`min_hessian`	1
`verbose`	1
`lambda_l1`	0.3
`device`	gpu
`num_threads`	8

Table 6: Hyperparameters for LightGBM.

The final average MSE obtained was 0.0289, and the average R-squared was 0.3865. This value is a relatively positive indicator, achieved while averaging the MSE of the serum triglyceride metric at 0.07. If serum triglycerides are excluded, the average MSE of the remaining four models is 0.0181, and the average R-squared is 0.3733. The figure of 0.0181 demonstrates that the model already achieves quite good performance.

R-squared is used to measure how much of the variation in the dependent variable can be explained by changes in the independent variables. However, medical diseases are often influenced by multiple factors, including genetics, environment, lifestyle, etc. This results in highly complex disease mechanisms, with factors that are difficult to account for or even impossible to explain, thereby limiting the model's explanatory power. Moreover, individual physiques cannot be generalized—different individuals may exhibit significant variations in their responses to and progression of diseases. Although the model's R-squared is only 0.3865, when considering the MSE, it can still be concluded that the model achieves relatively ideal results.

6.4 Analysis of Potential Pathogenic Factors

This study explores potential risk factors for *dual-high disease* through reverse feature analysis and identifies key medical examination items that significantly contribute to predicting *dual-high disease*. Here, the LightGBM model is employed to analyze high-contribution features, utilizing the model's `feature_importance()` method to obtain the importance scores of each feature.

The principle of this importance evaluation method operates through two key steps. First, the algorithm traverses each feature and attempts to split at every possible value of that feature. Second, it calculates the gain value at each split, which quantitatively represents the reduction in the loss function achieved by that particular split. These gain values serve as the fundamental metric for determining feature importance within the model.

Thus, each feature is assigned a corresponding gain value, where a higher gain value indicates greater feature importance.

Among the top 10 contributing features for diastolic and systolic blood pressure, 28 features are shared between both measurements. For diastolic blood pressure, systolic blood pressure, and serum triglycerides, there are 22 additional shared top 10 important features. The remaining two indicators have almost no overlapping features in their top 30 rankings.

Next, we manually examined these features for analysis. In the top 30 features for each indicator, most are anonymized numerical features whose meanings cannot be analyzed since the organizers have not disclosed their interpretations. Among the interpretable text features, we speculate the following correspondences: 0434 (Medical history), 4001 (Vascular function test), 1103 (Chest CT), 0113 (Liver), 0114 (Gallbladder), 0117 (Kidney), 0101 (Thyroid), 0115 (Pancreas), and 1001 (Electrocardiogram).

Feature 0434 ranks first for both diastolic and systolic blood pressure, likely representing medical history. Analysis suggests that individuals with histories of hypertension, diabetes, thyroid disorders, kidney disease, pancreatitis, fatty liver, obesity, or those undergoing related treatments are more susceptible to dual-high disease. Based on these findings, we had ChatGPT analyze relevant articles from NIH MedlinePlus and Mayo Clinic to identify underlying patterns.

Feature 4001 contains descriptors such as mild/moderate/severe reduction in vascular elasticity, atherosclerosis, and mild/moderate increase in arterial stiffness. This likely represents vascular function tests. Reduced vascular elasticity impairs normal blood-pumping function, atherosclerosis causes cholesterol plaque buildup in arterial walls, and increased arterial stiffness affects vasodilation/constriction capacity - all contributing to blood flow obstruction and elevated blood pressure.

Features 0113 (liver), 0114 (gallbladder), 0117 (kidney), 0101 (thyroid), and 0115 (pancreas) can be clearly inferred from their values. Fatty liver disease may lead to dual-high disease through mechanisms involving insulin resistance, abnormal lipid metabolism, inflammation, and oxidative stress. Insulin resistance causes fat accumulation in the liver, elevating blood triglyceride and cholesterol levels. Additionally, fatty liver-related inflammation and oxidative stress may exacerbate cardiovascular diseases, contributing to hypertension.

Regarding gallbladder function, impaired bile secretion/flow affects fat digestion, absorption of fat-soluble vitamins, and waste elimination. Impaired fat digestion leads to incomplete absorption, increasing blood lipid levels. Bile is crucial for absorbing fat-soluble vitamins; abnormal secretion reduces absorption, affecting normal metabolism. Cholesterol synthesized in the liver is excreted through bile; bile stasis or composition changes may reduce cholesterol excretion, increasing blood cholesterol levels. The liver and gallbladder mutually influence each other. Liver diseases like cirrhosis or hepatitis may affect bile production/flow, impacting gallbladder function. Conversely, gallbladder issues like gallstones or inflammation may trigger liver diseases, ultimately leading to abnormal blood lipids and pressure.

Kidney diseases may also contribute to dual-high disease. Conditions like nephritis impair glomerular filtration, compromising waste removal and water-salt balance regulation. This leads to fluid retention, increased blood volume, and consequently hypertension. The thyroid gland plays a vital role in the endocrine system. Hyperthyroidism increases metabolism and heart rate, raising cardiac workload and blood pressure. Hypothyroidism slows metabolism, potentially causing lipid metabolism disorders and elevated blood lipids.

Pancreatic disorders may cause insulin resistance and insufficient insulin secretion. As insulin regulates blood sugar, these conditions promote fat accumulation, elevating blood triglycerides and cholesterol levels. Electrocardiogram results can also indicate dual-high disease risk. Findings like tachycardia, bradycardia, or arrhythmia may increase hypertension risk by overactivating the sympathetic nervous system, releasing excess stress hormones (epinephrine/norepinephrine) that cause excessive vasoconstriction and increased heart rate. These conditions may also cause incomplete or rapid ventricular contractions that impair cardiac pumping, requiring increased force that elevates blood pressure.

Osteoporosis and hypertension show some correlation, possibly because osteoporosis is more prevalent in elderly populations who are more susceptible to dual-high disease. The renin-angiotensin-aldosterone system (RAAS) regulates blood pressure and affects bone metabolism. Angiotensin I and II influence osteoclast/osteoblast activity, with studies showing angiotensin II stimulates osteoclast proliferation. Hypertensive patients often exhibit increased intracellular calcium and renal calcium excretion. Excessive dietary sodium increases obligatory calcium excretion - a common potential factor for both conditions. Although the clinical relationship remains incompletely understood, studies show higher osteoporosis incidence among hypertensive patients compared to normal populations.

7 Acknowledgment

We acknowledge the use of AI-assisted tools in preparing this manuscript. While no content was directly generated by AI, we employed ChatGPT as a supplementary tool for specific linguistic purposes: (1) assisting with Chinese-to-English translation where our language proficiency was limited; (2) helping interpret Chinese reference materials; and (3) improving the fluency of certain passages. All substantive intellectual content, analysis, and conclusions remain our original work. Every AI-assisted portion underwent careful human verification to ensure academic integrity and accuracy.

References

G.V. Dhruva Kumar, V. Deepa, N. Vineela, G. Emmanuel, and Ch. Chittibabu. 2022. Lightgbm model based parkinson's disease detection by using spiral drawings. In *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 1–5.

J. B. Echouffo-Tcheugui, G. D. Batty, M. Kivimäki, and 1 others. 2013. Risk models to predict hypertension: a systematic review. *PLoS One*, 8(7):e67370.

B. Liao, X. Jia, T. Zhang, and 1 others. 2022. Dhdip: An interpretable model for hypertension and hyperlipidemia prediction based on emr data. *Computer Methods and Programs in Biomedicine*, 226:107088.

Yuming Qin. 2023. Research on hypertension prediction model based on xgboost algorithm. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, pages 809–812.

D. Sun, J. Liu, L. Xiao, and 1 others. 2017. Recent development of risk-prediction models for incident hypertension: An updated systematic review. *PLoS One*, 12(10):e0187240.

Udhaya T, Moulitharan M, Arockia Jegan S, and Akash M. 2025. Liver disease prediction using novel ensemble techniques with catboost and lightgbm. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pages 1388–1393.

Aiguo Wang, Ning An, Yu Xia, Lian Li, and Guilin Chen. 2014. A logistic regression and artificial neural network-based approach for chronic disease prediction: A case study of hypertension. In *2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*, pages 45–52.

Huazhong Yang, Zhongju Chen, Huajian Yang, and Maojin Tian. 2023. Predicting coronary heart disease using an improved lightgbm model: Performance analysis and comparison. *IEEE Access*, 11:23366–23380.

Jingyuan You, Juan Li, Xiaoyu Li, Haojie Li, Jinying Tu, Yuhuan Zhang, Jiandong Gao, Ji Wu, and Jingying Ye. 2023. Risk-prediction model for incident hypertension in patients with obstructive sleep apnea based on spo2 signals. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1–4.

Huanhuan Zhao, Zuchang Ma, and Yining Sun. 2019. A hypertension risk prediction model based on bp neural network. In *2019 International Conference on Networking and Network Applications (NaNA)*, pages 464–469.