Enhancing Lung Cancer Severity Prediction through Pattern Mining and Feature Engineering

Yuval Arbel Itamar Cohen

March 12, 2025

Abstract

Lung cancer severity prediction models often overlook complex interactions among diverse risk factors such as smoking, environmental exposures, and genetic predispositions. In this study, we integrate frequent pattern mining and association rule mining into the feature engineering stage of the data science pipeline for lung cancer datasets. Our approach systematically uncovers multi-factor patterns, transforms them into meaningful features, and incorporates them into a model (XGBoost). Experiments on four lung cancer datasets demonstrate that frequent pattern-based features can substantially improve accuracy and interpretability in some scenarios, particularly for multi-class severity classification. In one dataset, accuracy improved from roughly 36% to 60%. However, the effectiveness of this pattern-driven strategy varied across datasets, highlighting its dependence on dataset complexity and the presence of relevant factor interactions. This paper details all steps of data preprocessing, discretization, pattern discovery, grid-search tuning, and final model evaluation to illustrate a reproducible end-to-end pipeline. We conclude with insights for practical deployment and suggestions for future enhancements in mining-based feature engineering for clinical applications.

1 Introduction

Lung cancer remains one of the leading causes of cancer deaths worldwide. Although many machine learning algorithms have been proposed to predict lung cancer presence, severity, or survivability, they commonly rely on relatively straightforward features. Such features often do not fully capture the intricate, multi-factor etiology of the disease. Traditional single-feature or pairwise correlation methods can fail to detect more complex patterns involving interactions among multiple attributes (e.g., "chronic cough" *and* "exposure to air pollution" *and* "positive genetic risk").

To address this gap, we focus on improving the *pattern discovery* and *feature engineering* stages of the data science pipeline. We incorporate frequent pattern mining (FPM) and association rule mining (ARM) to uncover hidden structures in the data. These discovered patterns are subsequently

transformed into new features, which a supervised model can then exploit to achieve higher accuracy and potentially reveal clearer reasoning. This paper aims to provide a comprehensive account of how we preprocess medical data for pattern mining, and quantitatively compare our *Enhanced Model* to a standard *Baseline* across four datasets.

2 Related Work

Several prior studies have applied data mining methods to lung cancer. Agrawal and Choudhary [1] used association rule mining on SEER data to detect patient "hotspots," demonstrating the interpretability and domain value of discovered rules. Dimitoglou et al. [2] evaluated classical learning algorithms (Naïve Bayes, C4.5) for lung cancer survivability, highlighting the importance of robust feature selection and domain knowledge. Ahmed et al. [3] utilized an Apriori-based approach to identify frequent risk patterns for early detection, showcasing how multi-factor patterns can improve risk assessments.

Our work builds on these approaches by systematically embedding the discovered patterns into a predictive model, rather than only offering descriptive or stand-alone rule-based insights. Further, we test on four separate lung cancer datasets to illustrate generalizability and compare performance variations.

3 Datasets and Preprocessing

We consider four separate datasets (**D1**–**D4**), each containing different types of features (categorical, numeric, or ordinal) and targeting various outcomes (binary classification or multi-class severity). Table 1 provides an overview of each dataset's size, number of features, and target variable(s).

Table 1: Overview of the Four Lung Cancer Datasets

Dataset	**Samples**	**Features**	**Target Type**	**Example Target**
D1: Patients & Air Pollution	1000	24	3-class (severity)	`level` $\in \{0, 1, 2\}$
D2: Symptoms (Binary Task)	309	16	Binary (0/1)	`lung_cancer` $\in \{0, 1\}$
D3: Risk Factors & Stage	1008	37	Binary (0/1)	`stage` $\in \{0, 1\}$
D4: Patient Data (Binary)	999	16	Binary (0/1)	`lung_cancer` $\in \{0, 1\}$

3.1 Data Cleaning and Imputation

Prior to mining or modeling, we performed standard preprocessing steps:

1. **Missing Values**: For numerical features, missing values were filled using median imputation. Categorical features were first converted to strings and then label-encoded; no explicit mode imputation was applied.

2. **Outlier Handling**: Instead of explicitly removing or clipping extreme values (e.g., those beyond 4 standard deviations from the mean), numerical features were scaled using `StandardScaler` to normalize their distributions.

3. **Consistency Checks**: Rows with missing target values were dropped to ensure that only entries with valid labels were used. No additional checks for mislabeled entries were performed.

3.2 Discretization for Pattern Mining

Since frequent pattern and association rule mining generally work with categorical data, we discretized continuous variables into bins or categories where needed. For instance:

- `age` in D2 was bucketed into (18-40), (40-60) and (61+).
- `exposure_index` in D1 (a numeric measure of air pollution) was split into *low*, *medium*, *high* based on quartiles.
- Continuous lab results (like *blood_oxygen_level*) in D3 were binned by standard medical thresholds or quartiles.

After discretization, each distinct bin or category is treated as a possible "item" in the pattern mining step. We carefully chose bin boundaries using domain hints (e.g., `age`=60 as a known risk threshold) or data-driven quantiles to ensure balanced bin frequencies.

3.3 Train–Test Splits

To evaluate model performance fairly, each dataset was split into training and test sets of 70% training and 30% test. We stratified splits when possible (especially for binary tasks) to maintain similar class distributions in both splits.

4 Frequent Pattern and Association Rule Mining

We used the FP-Growth algorithm to discover frequent itemsets in the training portion of each dataset. We varied the *minimum support* threshold from 0.25, aiming to balance capturing meaningful patterns without drowning in too many low-frequency rules. Once frequent itemsets were identified, we generated association rules via standard metrics:

4.1 Rule Selection Criteria

- **Support**: how often an itemset occurs in the dataset. We set a lower bound of 25% for all the datasets.
- **Confidence**: P(Consequent | Antecedent). We typically used a 60% confidence minimum to consider a rule interesting enough.
- **Lift**: ratio of observed support to what would be expected if antecedent and consequent were independent. We looked for `lift` $= 1.0$ to highlight stronger-than-random associations.

Example Rule (in D1):

$$(\text{Smoking} \leq 2 \wedge \text{Chest Pain} \leq 2 \ \wedge \text{Yellow Fingers} \leq 2) \quad \Longrightarrow \quad (\text{level} = 2).$$

With support = 0.30, confidence = 0.75, lift = 1.58. his indicates that about 30% of the patients had {low smoking levels, chest pain, and yellow fingers}, and 75% of those were in severity level 2 (severe), making them 1.58 times more likely to be in that severity level than by random chance.

4.2 Feature Construction

We added the mined patterns as new binary (or numeric) features:

- **Binary Indicators**: For each discovered rule or pattern $\{i_1, i_2, \dots\} \implies$ Severe, we create a feature `pattern_A` which is 1 if the patient's profile has all $i_1, i_2, \dots$, else 0.
- **Composite Scores**: We also tested a simple additive scheme, e.g., `pattern_score`, counting how many risk patterns each patient meets.

If we discovered p patterns, we ended up with at most p new features (we took the top 5). However, we pruned patterns below thresholds of support/confidence to keep feature inflation manageable.

5 Modeling and Grid Search

We trained a **Baseline Model** on the original features only, and an **Enhanced Model** on the same features plus the new pattern-based features. We used the XGBoost framework for classification or regression (depending on the target). To ensure fair comparison, we performed the same hyperparameter grid search for each model, described below.

5.1 Hyperparameter Grid Search

For each dataset, we used 5-fold cross-validation on the training split to select the best hyperparameters. We explored:

- `n_estimators` $\in \{10, 50, 100\}$
- `max_depth` $\in \{1, 2, 3\}$
- `learning_rate` $\in \{0.01, 0.1, 1.0\}$
- `reg_alpha` (L1 penalty) $\in \{0, 10, 20\}$
- `reg_lambda` (L2 penalty) $\in \{0, 10, 20\}$

We chose hyperparameters that maximized validation performance (accuracy for classification, R^2 for regression). Then we retrained on the entire training set and evaluated on the hold-out test set.

Table 2: Sample of Selected Best Hyperparameters per Dataset

Dataset	`n_estim`	`max_depth`	`lrate`	`reg_alpha`	`reg_lambda`
D1	50	3	1.0	0	0
D2	50	3	0.1	0	20
D3	10	2	0.1	0	20
D4	50	3	0.01	0	10

6 Results and Discussion

We present both classification metrics (accuracy, precision, recall, F1) and regression metrics (RMSE, R^2) depending on the dataset.

6.1 Dataset D1: Multi-class Severity

Baseline results were unimpressive: overall accuracy around 0.36. Many samples in the test set ended up predicted as the same class. **Enhanced** with pattern-based features boosted accuracy to 0.60 (Table 3), a significant jump. Weighted F1 also increased substantially, from 0.19 to 0.48, indicating improved balance among the three severity classes. These pattern-based features appear especially beneficial in separating the most severe level (2) from mild or moderate.

6.2 Dataset D2: Binary Classification

D2 had relatively high baseline accuracy (about 0.87). The **Enhanced** model rose to 0.93 (Table 3), suggesting that frequent co-occurring symptoms and risk factors helped refine the decision boundary. Precision, recall, and F1 scores similarly improved (e.g., F1 from 0.81 to 0.89).

Table 3: Accuracy Comparison: Baseline vs. Enhanced

Dataset	**Baseline Accuracy**	**Enhanced Accuracy**
D1 (3-class)	0.363	0.603
D2 (binary)	0.871	0.925
D3 (binary/reg)	0.504	0.528
D4 (binary)	0.530	0.527

6.3 Dataset D3: Mix of Stage (binary)

For the classification portion (predicting `stage`), the baseline accuracy was around 0.50, and the enhanced approach reached 0.53. Although an improvement, it was modest. Examining confusion matrices suggests the model struggles to correctly classify stage=0; adding pattern features yields a small boost, but not large.

6.4 Dataset D4: Binary Classification

We observed negligible change in accuracy: from 0.53 to 0.527. In-depth analysis of the final patterns showed that many discovered itemsets overlapped heavily with existing features. Thus, the model did not gain new signal from them. Some pattern features even correlated highly with original ones, introducing partial redundancy.

6.5 Interpretability and Rule Examples

One notable advantage of pattern-based augmentation is interpretability. For example, a discovered rule in D1 might read:

$$(\text{Passive Smoke} \leq 4.0 \wedge \text{Dry Cough} \leq 4.0) \Longrightarrow \text{level}=2$$

This suggests that about 61% of the patients had both {Passive Smoker, Dry Cough}, and a significant portion of them belonged to severity level 2 (severe)

7 Analysis of Findings

Overall, the **Enhanced Model** outperformed the **Baseline** significantly in two datasets (D1, D2). In D3, the improvement was modest for classification and moderate for regression. In D4, no net benefit was observed. These results highlight key lessons:

1. **Data Richness and Redundancy:** If existing features or domain-specific attributes already capture the relationships, pattern mining adds less.
2. **Relevance of Mined Patterns:** Setting appropriate support/confidence thresholds is critical. Overly strict thresholds can miss valuable interactions; overly lenient thresholds can introduce noisy or redundant features.
3. **Class Distribution and Task Difficulty:** In multi-class scenarios (like D1), pattern-based features can help separate overlapping classes if meaningful factor combinations exist.

8 Limitations and Future Work

While promising, our method has some drawbacks:

- **Scalability:** Frequent pattern mining can become expensive for large-scale or high-dimensional data if not carefully pruned.
- **Feature Explosion:** Each discovered pattern can become a new feature, risking overfitting or extremely large input dimensions.
- **Threshold Sensitivity:** The support/confidence cutoffs can drastically alter the set of patterns, impacting results.

Potential extensions:

1. Use **sequence mining** if temporal data (e.g., disease progression timeline) is available, capturing the order of factor occurrences.
2. **Domain-Guided Pruning:** Incorporate medical knowledge to filter or prioritize patterns, improving both interpretability and performance.
3. **Advanced Feature Selection**: After generating new features, apply methods such as L1 regularization or mutual information to remove redundant pattern-based features.
4. **Further Model Experimentation**: Evaluate whether neural networks or other ensemble methods might exploit pattern-based features differently.

9 Conclusion

In this paper, we demonstrated a pattern-mining-driven approach to enhance lung cancer severity prediction, integrating frequent itemset discovery and association rule mining into the data science pipeline. Our comprehensive experiments across four distinct datasets revealed substantial accuracy gains (e.g., +24% on D1, +5% on D2) when the mined patterns were novel and relevant, while certain cases (D4) saw negligible improvement due to feature redundancy. Moreover, the discovered patterns offer interpretable, multi-factor insights (e.g., a triad of symptoms or risk factors collectively signaling severity), underscoring their practical value in medical contexts.

In conclusion, while pattern-based features are not universally beneficial, they show strong potential for capturing intricate lung cancer risk interactions. Future directions revolve around refining pattern selection, improving interpretability, and addressing scalability. By bridging data mining and predictive modeling, this study highlights a flexible methodology that can be adapted to a range of clinical or other complex, multi-factor domains.

References

[1] A. Agrawal and A. Choudhary, Identifying HotSpots in Lung Cancer Data Using Association Rule Mining. In: 2011 IEEE Int. Conf. on Data Mining Workshops (ICDMW), pp. 995–1002, IEEE, 2011.

[2] G. Dimitoglou, J. A. Adams, and C. M. Jim, Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability. arXiv:1206.1121, 2012.

[3] K. Ahmed, A. Al-Emran, T. Jesmin, R. F. Mukti, M. Z. Rahman, and F. Ahmed, Early Detection of Lung Cancer Risk Using Data Mining. Asian Pacific Journal of Cancer Prevention, 14(1): 595–598, 2013.