Predictive Modeling of Lung Cancer Risk in Workers Using Dropouts Meet Multiple Additive Regression Trees

Haewon Byeon
Department of Future Technology, Korea University of Technology and Education (KOREA TECH),
Cheonan 31253, South Korea

***Abstract*—Lung cancer remains a leading cause of cancer mortality, and preventable occupational and environmental exposures may compound risk in working-age populations. This study developed and compared predictive models for lung cancer risk using a publicly available tabular dataset (Kaggle; n = 1,000) containing demographic, lifestyle, symptom, and exposure-related variables. After standard preprocessing and an 80/20 train-test split, a Classification and Regression Tree (CART), a dropout-regularized gradient-boosted tree model (DART), k-nearest neighbors (KNN), and Gaussian Naïve Bayes were trained and evaluated using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). CART achieved the highest accuracy (84.5%), while KNN achieved the highest precision (78.7%). DART produced the best F1-score (77.3%) and the highest AUC (0.801), suggesting a favorable balance between sensitivity and specificity when accounting for class imbalance. Feature-importance patterns in the final DART model highlighted occupational hazards, smoking habits, genetic predisposition, and air pollution exposure as leading contributors to model-based risk stratification in occupational settings. These findings suggest that regularized ensemble tree methods can support stable risk stratification and may complement screening by prioritizing individuals who warrant closer evaluation. The analysis is limited by the modest sample size and reliance on a single public dataset; external validation in occupational cohorts with measured exposure histories is required before practical implementation.**

Keywords*—Lung cancer; predictive modeling; occupational exposure; gradient boosting; risk stratification

I. Introduction

Lung cancer remains a major driver of cancer morbidity and mortality worldwide [1]. Within this global burden, the occupational sector represents a critical demographic where work-related exposures add substantially to baseline risk [2]. The etiology of lung cancer in workers is rarely driven by a single factor; instead, ambient air pollution and workplace hazards often co-occur and may act jointly in lung carcinogenesis [3,4]. Furthermore, individual lifestyle factors, such as smoking and dietary patterns, further shape susceptibility [5].

Despite these complex interactions, current occupational health programs often rely on broad exposure categories or rigid age-based screening criteria. Consequently, workers with nuanced risk profiles-such as those combining moderate air pollution exposure, intermittent contact with industrial dusts, and long-term smoking-are easily missed in routine checkups. Taken together, these overlapping determinants motivate the need for advanced models that can identify high-risk workers early, when prevention and screening have the greatest leverage. A prediction model that aggregates these multifaceted signals can serve as crucial decision support, helping to prioritize counseling, exposure reduction, or clinical follow-up.

Predictive modeling offers a solution by translating observed risk-factor patterns into estimates of future disease outcomes [6,7]. While traditional statistical methods have limitations in capturing high-dimensional data, machine-learning approaches are attractive because they accommodate nonlinear relationships and complex interactions between demographic characteristics, exposure histories, and health behaviors.

This study evaluates Dropouts meet Multiple Additive Regression Trees (DART) for lung cancer risk stratification among workers and compares its performance with Classification and Regression Trees (CART), Naïve Bayes, and K-Nearest Neighbors (KNN). Performance is assessed using accuracy, precision, recall, F1-score, balanced accuracy, and area under the curve (AUC). The top five predictors in the final DART model are also ranked to clarify which factors drive risk stratification in this dataset. The novelty of the study lies in the systematic evaluation of dropout-regularized boosting in an occupationally relevant tabular setting that combines exposure-related, behavioral, and symptom variables. Unlike prior studies that mainly emphasize general clinical or imaging-based prediction, the present analysis focuses on interpretable screening-oriented risk grouping and compares DART with widely used baseline classifiers using class-sensitive performance metrics.

Section II reviews related literature, Section III describes the methods, Section IV reports the results, and Section V discusses implications and future research directions. Finally, Section VI concludes the study.

II. Literature Review

Predictive modeling has become a routine component of modern clinical and public health research, fueled by the rapid growth of available health data and advances in machine-learning methods [8-10]. Lung cancer prediction has drawn particular attention because early detection remains difficult, and risk is influenced by multiple domains of exposure and

behavior [11].

Lung cancer etiology is multifactorial, spanning genetic susceptibility, environmental exposures, and lifestyle factors. Traditional epidemiologic models remain useful, but they often struggle with nonlinear associations and high-order interactions when many predictors are considered simultaneously [12-14]. Machine-learning algorithms such as support vector machines (SVMs), random forests, and neural networks can model these complexities and have often outperformed conventional approaches in prediction settings [15,16].

Prior studies suggest that machine-learning methods can improve diagnostic and prognostic performance in lung cancer applications [17,18]. Random forests, for example, are frequently used because they tolerate high-dimensional data and reduce variance through ensembling; several reports have used them to estimate lung cancer risk with strong performance [19]. SVM-based classifiers have also been applied to distinguish cancerous from non-cancerous cases, benefiting from their ability to operate in high-dimensional feature spaces [20]. When paired with feature selection, these models can focus learning on the most informative variables and reduce noise.

Occupational exposures remain a central theme in lung cancer research. Workers in mining, construction, and manufacturing may experience sustained contact with carcinogens such as asbestos, silica, and diesel exhaust, which can accumulate risk over time. Studies that incorporate exposure indicators alongside behavioral variables generally report better discrimination than models that rely on demographics alone.

Overall, prior work indicates that data-driven models can improve lung cancer risk prediction when they integrate heterogeneous information sources. These models can also help prioritize which exposures and behaviors warrant intervention. Building on this literature, DART is applied here to a worker-relevant risk-factor dataset, and its performance is assessed against commonly used baseline classifiers.

Recent work has also shifted from single-model comparisons to ensemble strategies that improve stability. Tree-based boosting methods are attractive because they handle mixed data types and often perform well with limited feature engineering [15,16]. Standard boosting can overfit when a small set of trees dominates successive updates. DART mitigates this behavior by randomly dropping trees during training, encouraging the ensemble to distribute weight across multiple components. The mechanism resembles dropout in neural networks and can improve generalization while preserving an additive structure that remains interpretable.

However, prior lung cancer prediction studies still leave several gaps. Many studies focus on general clinical or imaging-based prediction and give limited attention to occupationally relevant exposure indicators within tabular risk-stratification settings. In addition, several reports emphasize overall accuracy more than class-balanced discrimination, although screening decisions depend on the trade-off between missed high-risk cases and unnecessary follow-up. Evidence comparing dropout-regularized boosting with simpler baseline classifiers in this context also remains limited. To address these gaps, the present study evaluates DART against CART, KNN, and Naïve Bayes using exposure-related, behavioral, and symptom variables, with emphasis on AUC, balanced accuracy, and F1-score for occupationally relevant risk stratification.

III. Methodology

A. Dataset Description

The Kaggle "Lung Cancer Prediction" dataset (https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link?resource=download) was used because it includes demographic, environmental, lifestyle, and symptom-related variables relevant to lung cancer risk. The dataset contains predictor variables and a categorical outcome ("Level of cancer") with three classes (low, medium, high). Table I summarizes each feature and its coding.

Several variables in Table I represent symptoms (e.g., cough, chest pain, fatigue) in addition to exposures and behaviors. The outcome label is ordinal (low/medium/high), but it is treated here as a three-class classification problem because the dataset does not provide time-to-event information. This framing matches common screening workflows where clinicians must distinguish relative risk groups rather than predict incidence.

TABLE. I. Variables Included in the Lung Cancer Prediction Dataset and their Coding Scheme.

Feature	Description	Classes
Age	Numeric representation of the patient's age.	-
Gender	Categorical variable indicating the patient's gender.	Male, Female
Air Pollution	Categorical measure of the patient's exposure to air pollution.	Low, Medium, High
Alcohol Use	Categorical level of alcohol consumption.	None, Low, Moderate, High
Dust Allergy	Categorical level indicating susceptibility to dust allergies.	None, Mild, Severe
Occupational Hazards	Categorical level of exposure to occupational risks.	None, Low, Medium, High
Genetic Risk	Categorical assessment of genetic predisposition to lung cancer.	Low, Medium, High
Chronic Lung Disease	Categorical measure of chronic lung disease presence.	None, Mild, Severe
Balanced Diet	Categorical level of diet balance.	Poor, Average, Good
Obesity	Categorical level indicating obesity.	Underweight, Normal, Overweight, Obese
Smoking	Categorical measure of smoking habits.	Non-smoker, Occasional, Regular
Passive Smoker	Categorical level of exposure to secondhand smoke.	None, Low, High
Chest Pain	Categorical level of chest pain experienced.	None, Occasional, Frequent
Coughing of Blood	Categorical level indicating coughing of blood.	None, Rare, Frequent
Fatigue	Categorical level of fatigue.	None, Mild, Severe
Weight Loss	Categorical level of weight loss.	None, Mild, Severe
Shortness of	Categorical level indicating	None, Mild, Severe

Breath	shortness of breath.	
Wheezing	Categorical level of wheezing.	None, Occasional, Frequent
Swallowing Difficulty	Categorical measure of difficulty in swallowing.	None, Mild, Severe
Clubbing of Finger Nails	Categorical level of finger clubbing.	None, Mild, Severe
Frequent Cold	Categorical frequency of colds.	Rare, Occasional, Frequent
Dry Cough	Categorical level of dry cough.	None, Mild, Severe
Snoring	Categorical level of snoring.	None, Mild, Severe
Level of cancer	Target variable indicating the risk level of lung cancer.	Low, Medium, High

B. Preprocessing Techniques

1) Data cleaning and missing value imputation: The preprocessing pipeline begins with basic data cleaning to address inconsistencies, invalid entries, and missing values. The data were screened for out-of-range values, category labels were harmonized, and obvious duplicates were removed where applicable. Because missingness can bias training and evaluation, standard imputation methods were applied to produce a complete analysis dataset.

For numeric variables, mean imputation replaces each missing value with the arithmetic mean of observed values for that feature. This approach is expressed mathematically as:

Mean imputation replaces missing numeric values with the arithmetic mean of the observed values (sum of observed values divided by n), where x_i represents the data point, and n is the number of non-missing observations.

For categorical variables, mode imputation was used, with missing entries replaced by the most frequently observed category. This choice preserves the observed distribution of categories and is simple to implement in small to moderate datasets.

2) Encoding categorical variables: Several predictors are categorical and must be converted to numeric representations before model fitting. One-hot encoding was used to represent each category as a separate binary indicator, which avoids imposing an artificial ordering on nominal categories.

One-hot encoding maps a categorical variable C with categories {c1, ..., ck} to k binary indicators, where the indicator for category c_j is 1 when C equals c_j and 0 otherwise.

This representation allows the model to learn from categorical features while treating categories as distinct states rather than as ordered values.

3) Feature scaling and normalization: Feature scaling was also applied where appropriate. Although tree-based methods are typically robust to monotonic transformations, scaling can stabilize distance-based models such as KNN and improve numerical behavior during model comparison.

Standardization (Z-score normalization) rescales a feature to have mean 0 and standard deviation 1. The formula for standardization is:

Z-score standardization transforms a feature as $z = (x - \mu) / \sigma$, where x is the original value, μ is the feature mean, and σ is the standard deviation.

Normalization rescales a feature to a fixed interval, typically [0, 1], using the formula:

Min-max scaling rescales a feature as $x' = (x - \text{minx}) / (\text{maxx} - \text{minx})$, where minx and maxx are the feature's minimum and maximum.

Together, these steps yield a dataset that is consistently encoded and suitable for training multiple model classes under comparable conditions.

C. Modeling Techniques

The primary model in this study is DART, an extension of gradient boosting that introduces a dropout mechanism during boosting. By randomly dropping trees when updating the ensemble, DART reduces the tendency of boosted models to overfit and can improve generalization while preserving the strengths of additive tree ensembles.

For comparison, three widely used baseline classifiers were trained: CART as an interpretable single-tree model, Naïve Bayes as a probabilistic benchmark under conditional independence assumptions, and KNN as a nonparametric method based on local neighborhood structure.

Models were evaluated using accuracy, precision, recall, and F1-score, and balanced accuracy and AUC were also reported to account for class imbalance and threshold-independent discrimination. This evaluation choice follows established recommendations for assessing prediction model performance and reflects the known sensitivity of nominal accuracy to skewed class distributions in screening-style datasets [21-25]. These metrics provide complementary views of performance in risk prediction settings.

Accuracy alone can hide systematic errors when classes are imbalanced or when one class is easier to identify than the others. Balanced accuracy averages recall across classes so that each class contributes equally. AUC summarizes rank-order discrimination across thresholds, and ROC analysis was interpreted according to standard methodological guidance [23,24]. In multiclass settings, AUC is commonly computed by averaging one-vs.-rest curves. Precision and recall quantify the trade-off between false alarms and missed cases, which is central in screening. Reporting these metrics together clarifies whether performance gains come from better detection, fewer false positives, or both.

D. Experimental Setup

1) Data splitting and validation techniques: The data were split into training and test sets using an 80/20 stratified partition, preserving the class distribution of the "Level of cancer" outcome. The training set was used for model fitting and tuning, and the test set provided an out-of-sample assessment of performance.

k-fold cross-validation was also applied within the training data to reduce variance in performance estimates and to support hyperparameter selection. This procedure rotates validation

folds across the training set, limiting the dependence of results on any single split.

2) Parameter settings for models: Table II lists the parameter settings used for each model. Values were selected to balance performance and interpretability, and the baselines were kept simple to provide a clear reference point.

TABLE. II. HYPERPARAMETER SETTINGS USED FOR DART AND BASELINE MODELS

Model	Parameters
DART	Learning Rate: 0.1, Number of Trees: 100, Dropout Rate: 0.2
CART	Max Depth: 5, Min Samples Split: 10
Naïve Bayes	Prior Probabilities: Estimated from Data
K-Nearest Neighbors	Number of Neighbors (K): 5, Distance Metric: Euclidean

For DART, the learning rate and number of trees were tuned to manage the bias-variance trade-off, and a dropout rate was used to reduce overfitting by omitting trees during training. CART hyperparameters were set to limit tree depth and avoid overly complex partitions. Naïve Bayes required minimal tuning beyond estimating class priors from the data. For KNN, K = 5 was used based on empirical testing that balanced sensitivity to noise with decision-boundary flexibility.

IV. RESULTS AND ANALYSIS

A. Comparative Model Performance

All models were trained on the same preprocessed feature set and evaluated on the held-out test split to keep comparisons fair. The results are interpreted with occupational screening in mind, where the cost of missing high-risk workers may differ from the cost of recommending follow-up for lower-risk individuals. For that reason, accuracy is discussed alongside metrics that reflect class-wise sensitivity and threshold-independent discrimination.

Pairwise correlations among predictors were examined to assess redundancy before modeling. The correlation matrix showed moderate associations for several feature pairs, but no correlation coefficient exceeded 0.85. This pattern suggests limited multicollinearity and supports retaining the full feature set for model training. Fig. 1 presents the feature correlation matrix used for this assessment.

TABLE. III. TEST-SET PERFORMANCE ACROSS MODELS (ACCURACY, PRECISION, RECALL, AND F1-SCORE).

Model	Accuracy	Precision	Recall	F1-Score
DART	0.787	0.716	0.740	0.773
CART	0.845	0.716	0.751	0.711
Naïve Bayes	0.823	0.706	0.682	0.708
KNN	0.810	0.787	0.777	0.708

Performance differed across models in ways that matter for occupational risk screening (Table III). CART achieved the highest overall accuracy (84.5%), indicating strong classification performance on this dataset. DART achieved an accuracy of 78.7% and showed more balanced behavior across several metrics, which may be useful when misclassification costs are asymmetric.

Precision and recall capture different error profiles. KNN achieved the highest precision (78.7%), while CART showed the highest recall (75.1%). DART produced the highest F1-score (77.3%), indicating a favorable balance between precision and recall in this setting.

Compared with the baseline classifiers, DART did not yield the highest nominal accuracy, but it showed the most favorable balance for screening-oriented prediction by achieving the highest F1-score and AUC. This pattern indicates that DART better preserves discrimination across risk levels while reducing over-reliance on a single dominant decision boundary. In occupational screening, where missed high-risk cases may be more consequential than a modest increase in false positives, this balanced performance represents a practical advantage over CART, KNN, and Naïve Bayes.

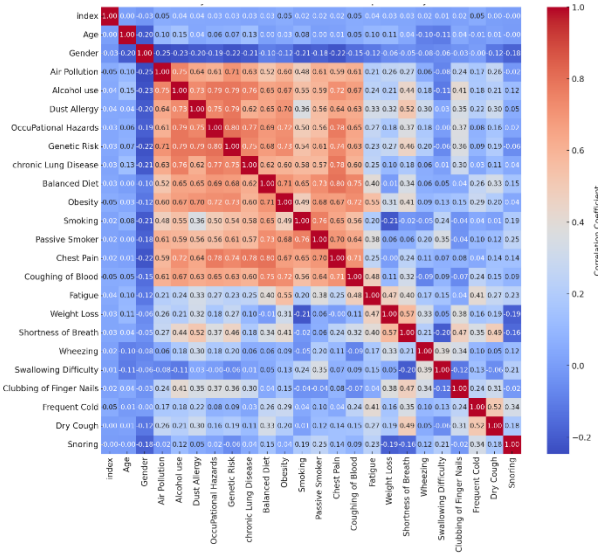

Fig. 1. Feature correlation matrix for variables in the lung cancer prediction dataset.

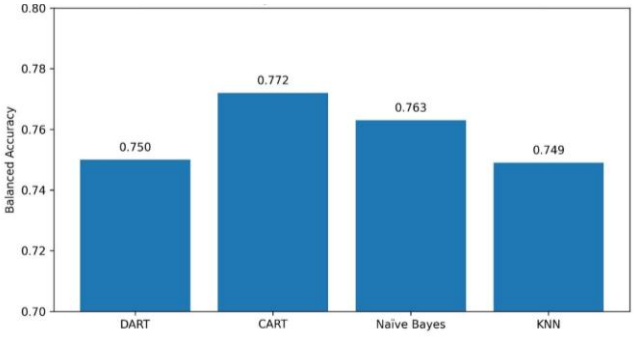

Fig. 2. Balanced accuracy across models on the held-out test set.

Balanced accuracy and AUC offer additional perspective beyond accuracy alone. DART achieved the highest AUC (80.1%), suggesting strong rank-order discrimination across risk levels. Its balanced accuracy was competitive with the other models, supporting its use when class proportions are uneven and a single accuracy figure can be misleading.

The "Level of cancer" outcome spans three classes, and class proportions can inflate naive accuracy if a model favors the most common class. Because this imbalance problem is well documented in classification research [21,22], balanced accuracy helps guard against this failure mode, and Fig. 2 provides a visual comparison that is easier to scan.

Beyond overall performance, the predictors that most strongly influenced the final DART model were examined. The feature-importance rankings highlighted occupational hazards, smoking habits, genetic predisposition, and air pollution as leading contributors, consistent with established risk pathways. These rankings help translate the model output into actionable targets for workplace and behavioral interventions.

Across metrics, DART performed competitively while providing a clear feature-importance profile. Combined, these results suggest that DART can support risk stratification in occupational health contexts, particularly when the goal is to identify high-risk groups for targeted prevention and follow-up.

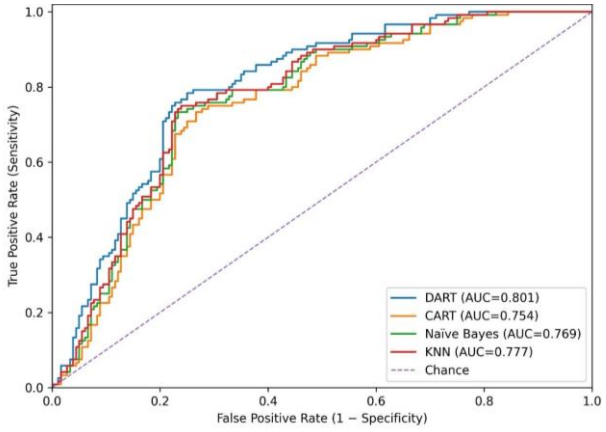

Fig. 3. Area under the ROC curve (AUC) across models on the held-out test set.

B. Feature Importance Analysis

Because DART is an ensemble of trees, it provides natural feature-importance measures based on the contribution of each variable to splits that reduce prediction error. The top five features are reported to emphasize factors that dominate model decisions in this dataset.

Occupational Hazards ranked highest, which aligns with epidemiologic evidence linking long-term workplace exposure to lung cancer. This result supports routine exposure assessment and stronger controls in high-risk industries.

Importance scores summarize how much each variable contributed to error reduction across the ensemble. Higher scores indicate features that the model repeatedly used to separate risk levels during training.

Smoking Habits ranked second. The model treated smoking as a dominant modifiable risk factor, reinforcing the value of cessation support as part of workplace health programs.

Genetic Predisposition also contributed meaningfully. While genetics are not modifiable, identifying workers with family history or related indicators may help guide tailored screening or counseling.

Exposure to Air Pollution showed substantial importance, particularly relevant for urban and industrial settings. Reducing ambient exposure may require both workplace measures and broader environmental policy.

Balanced Diet had a smaller importance score (0.05), suggesting that diet is a secondary contributor in this dataset. Even so, diet may interact with other behaviors and exposures, and it remains a reasonable target for health promotion.

Feature-importance measures in tree ensembles are useful but not definitive. They reflect how variables reduce prediction error, which can be influenced by correlation among predictors and by the coding of categorical levels. A high importance score does not imply causality, and a low score does not rule out clinical relevance. Table IV is interpreted as a prioritization tool for further epidemiologic study and targeted intervention planning, not as causal evidence.

TABLE. IV. TOP FIVE PREDICTORS AND THEIR IMPORTANCE SCORES IN THE FINAL DART MODEL.

Feature	Importance Score
Occupational Hazards	0.25
Smoking Habits	0.22
Genetic Predisposition	0.18
Air Pollution Exposure	0.15
Balanced Diet	0.05

V. DISCUSSION

This study demonstrated that the DART model serves as a robust and practical tool for predicting lung cancer risk among workers. While the CART algorithm achieved the highest overall accuracy, accuracy alone can be a misleading metric in imbalanced datasets typical of cancer screening. In contrast, DART exhibited superior discriminative ability (AUC 0.801) and a balanced F1-score (Fig. 3). This balance is critical in occupational health settings, where the cost of a false negative (missing a cancer case) is significantly higher than that of a false positive. By effectively handling the trade-off between precision and recall, DART provides a more reliable safety net for identifying high-risk individuals who warrant further clinical investigation.

Beyond predictive performance, the feature importance analysis aligns the model's "black box" predictions with established biological plausibility. The high ranking of occupational hazards and smoking history is consistent with prior epidemiological evidence, confirming that the model is detecting genuine signal rather than noise. Furthermore, the model captured significant contributions from genetic predisposition, suggesting that family history remains a vital component of risk assessment. Crucially, machine-learning approaches like DART can synthesize multiple risk domains-demographic, symptomatic, and environmental-into a single framework. This integration is particularly valuable in modern workplaces where workers face layered risks, such as the combined effect of moderate air pollution and industrial dusts, rather than a single dominant exposure.

From a practical standpoint, deploying a data-driven risk score could transform routine occupational health surveillance. Instead of relying solely on age or job title, employers and occupational physicians could use these risk stratifications to prioritize resources. High-risk scores triggered during periodic health examinations could justify targeted interventions, such as smoking-cessation support, stricter exposure monitoring, or referral for low-dose CT screening. Policymakers may also leverage these findings to identify high-risk sectors that require tighter exposure limits. However, the operating threshold for these interventions must be calibrated to local resources and baseline exposure profiles to balance clinical benefit against the risk of over-testing.

Several limitations should be considered when interpreting these results. First, the analysis relies on a single public dataset, which may not capture the full diversity of occupational exposures across different industries or countries, limiting generalizability. Second, while feature importance estimates aid interpretation, they represent statistical associations and do not establish causal mechanisms for lung carcinogenesis. Third, complex ensemble models like DART present interpretability challenges compared to simple decision trees; future implementation will require "Explainable AI" XAI methods to build clinical trust [26, 27]. Finally, unmeasured confounding and dataset biases may remain.

Future work should validate these findings in external cohorts with richer exposure histories and longitudinal follow-up. Such data would support time-to-event modeling, moving beyond cross-sectional classification to incident lung cancer prediction. Additionally, subgroup analyses by industry and age are necessary to ensure equitable deployment across diverse worker populations. To improve reproducibility and transparency, future studies should also report model development and validation in line with TRIPOD guidance [26]. Continued collaboration among occupational physicians, epidemiologists, and data scientists will be essential to translate these predictive models into standard preventive practice.

VI. Conclusion

This study demonstrates that the Dropouts meet Multiple Additive Regression Trees (DART) algorithm offers a robust approach for stratifying lung cancer risk among workers, particularly when evaluation is based on class-sensitive performance rather than nominal accuracy alone. While CART achieved the highest overall accuracy, DART provided the highest AUC (0.801) and F1-score (0.773), indicating a more balanced trade-off between discrimination and screening-oriented error control. Feature-importance analysis highlighted occupational hazards, smoking habits, genetic predisposition, and air pollution exposure as leading contributors to model-based risk stratification. These findings suggest that data-driven risk scores may support occupational health surveillance by helping prioritize preventive counseling and clinical follow-up for high-risk groups. These findings should be interpreted cautiously because the analysis was based on a relatively small single public dataset, and several variables may reflect self-reported or proxy information that is susceptible to measurement bias. External validation using larger occupational cohorts remains necessary before practical implementation. Future studies should validate the model in independent occupational cohorts, incorporate measured exposure histories and longitudinal follow-up, and examine model performance across industries and age groups. Ultimately, this line of research supports a shift from reactive diagnosis toward more proactive and targeted prevention in occupational health.

Acknowledgment

This research through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF- RS-2023 00237287).

References

[1] D. R. Aberle et al., “Reduced lung-cancer mortality with low-dose computed tomographic screening,” New England Journal of Medicine, vol. 365, no. 5, pp. 395-409, 2011.

[2] H. J. de Koning et al., “Reduced lung-cancer mortality with volume CT screening in a randomized trial,” New England Journal of Medicine, vol. 382, no. 6, pp. 503-513, 2020.

[3] US Preventive Services Task Force, “Screening for lung cancer: US Preventive Services Task Force recommendation statement,” JAMA, vol. 325, no. 10, pp. 962-970, 2021.

[4] M. C. Tammemägi et al., “Selection criteria for lung-cancer screening,” New England Journal of Medicine, vol. 368, no. 8, pp. 728-736, 2013.

[5] S. A. Kovalchik et al., “Targeting of low-dose CT screening according to the risk of lung-cancer death,” New England Journal of Medicine, vol. 369, no. 3, pp. 245-254, 2013.

[6] P. B. Bach et al., “Variations in lung cancer risk among smokers,” Journal of the National Cancer Institute, vol. 95, no. 6, pp. 470-478, 2003.

[7] M. R. Spitz et al., “A risk model for prediction of lung cancer,” Journal of the National Cancer Institute, vol. 99, no. 9, pp. 715-726, 2007.

[8] A. Cassidy et al., “The LLP risk model: An individual risk prediction model for lung cancer,” British Journal of Cancer, vol. 98, no. 2, pp. 270-276, 2008.

[9] H. A. Katki et al., “Development and validation of risk models to select ever-smokers for computed tomography lung cancer screening,” JAMA, vol. 315, no. 21, pp. 2300-2311, 2016.

[10] H. A. Katki et al., “Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening,” Annals of Internal Medicine, vol. 169, no. 1, pp. 10-19, 2018.

[11] A. McWilliams et al., “Probability of cancer in pulmonary nodules detected on first screening CT,” New England Journal of Medicine, vol. 369, no. 10, pp. 910-919, 2013.

[12] D. Ardila et al., “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” Nature Medicine, vol. 25, no. 6, pp. 954-961, 2019.

[13] A. Hosny et al., “Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study,” PLoS Medicine, 15(11), e1002711, 2018.

[14] H. J. W. L. Aerts et al., “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” Nature Communications, 5, Article 4006, 2014.

[15] P. Lambin et al., “Radiomics: The bridge between medical imaging and personalized medicine,” Nature Reviews Clinical Oncology, vol. 14, no. 12, pp. 749-762, 2017.

[16] Breiman and L, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[17] C. Cortes, and Vapnik, and V, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[18] J. H Friedman, “Greedy function approximation: A gradient boosting machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[19] Tibshirani and R, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[20] I. Guyon, and Elisseeff, and A, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P and Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[22] H. He and E. A and Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[23] J. A. Hanley and B. J and McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29-36, 1982.

[24] Fawcett and T, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[25] E. W. Steyerberg et al., "Assessing the performance of prediction models: A framework for traditional and novel measures," Epidemiology, vol. 21, no. 1, pp. 128-138, 2010.

[26] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M and Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," Annals of Internal Medicine, vol. 162, no. 1, pp. 55-63, 2015.

[27] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.