Norwegian University of Science
and Technology

TDT4259 - Applied Data Science

Group Assignment

A Data-Driven Approach to Lifestyle Risk Prediction Using Clinical Biomarkers: Smoking and Drinking Analysis

---

Authors:

Marco Prosperi - 151613
Andrea Richichi - 151790
Tizita Belachew Tamirat - 128549
Gianluigi Vazzoler - 152698

Fall, 2025

Contents

1 Introduction

1.1 Background and Context

Lifestyle-related health risks, particularly smoking and alcohol consumption, represent significant public health challenges worldwide. These behaviors are associated with numerous chronic diseases, including cardiovascular disease, cancer, liver disease, and respiratory conditions. Traditional approaches to identifying at-risk individuals often rely on self-reported surveys, which can be subject to underreporting due to social desirability bias or lack of awareness.

Recent advances in preventive medicine have highlighted the potential of using objective biomedical data to identify individuals with risky lifestyles before serious health complications arise. Clinical biomarkers—measurable indicators derived from routine health examinations—provide objective evidence of physiological changes associated with smoking and drinking behaviors. These markers include liver function enzymes (AST, ALT), cholesterol levels, blood pressure measurements, and anthropometric data.

The ability to predict lifestyle risks from biomedical data has important implications for healthcare providers. Early identification of at-risk individuals enables targeted interventions, personalized health counseling, and preventive care strategies. This proactive approach can reduce the burden of lifestyle-related diseases on healthcare systems and improve patient outcomes.

1.2 Problem Definition

Despite the availability of extensive biomedical data from routine health screenings, healthcare providers often lack effective tools to systematically identify individuals with risky lifestyle behaviors. The challenge lies in developing predictive models that can accurately classify individuals based on smoking and alcohol consumption patterns using objective clinical measurements, provide interpretable predictions that healthcare professionals can act upon, and identify the most relevant biomarkers that signal lifestyle-related health risks.

The primary research question addressed in this project is: *can machine learning models effectively predict smoking status and alcohol consumption levels using clinical biomarkers, and which biomarkers are most predictive of these lifestyle behaviors?*

1.3 Motivation

The motivation for this project stems from multiple converging factors in modern healthcare. The global burden of lifestyle-related diseases continues to rise, with smoking and excessive alcohol consumption being among the leading preventable causes of mortality. According to the World Health Organization, tobacco use kills more than 8 million people annually, while harmful use of alcohol results in approximately 3 million deaths each year.

Traditional screening methods based on self-reporting have well-documented limitations in accuracy and reliability. Individuals may underreport their consumption due to stigma, denial, or simple forgetfulness. An objective, data-driven approach using readily available biomedical measurements could overcome these limitations and provide healthcare providers with unbiased risk assessments.

Furthermore, routine health examinations already collect extensive biomedical data, but this information is often underutilized for predictive purposes. By developing machine learning models that leverage this existing data, healthcare providers can implement risk prediction systems without requiring additional invasive tests or costly procedures. This represents an opportunity to extract more value from data that is already being collected.

Early identification of at-risk individuals creates opportunities for preventive interventions that are more effective and less costly than treating advanced disease. This aligns with the

broader shift in healthcare toward preventive and personalized medicine, where the focus moves from reactive treatment to proactive risk management.

1.4 Team and Roles

This project was developed by master's students from the Norwegian University of Science and Technology (NTNU) as part of the Applied Data Science course.

Team Member	Student ID	Primary Contributions
Marco Prosperi	151613	Technical documentation and report writing
Andrea Richichi	151790	Presentation and video production
Tizita Belachew Tamirat	128549	Project planning and report writing
Gianluigi Vazzoler	152698	ML pipeline implementation, model development

Table 1: Team members and their primary contributions

1.5 Report Outline

This report is structured to provide a comprehensive overview of the predictive modeling approach for lifestyle risk identification. The organization follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring systematic progression from problem understanding to deployment recommendations.

Section 2 defines the specific goals of the analysis, project limitations, and the broader context within preventive healthcare. Section 3 describes the dataset sources, preprocessing procedures, quality assurance measures, and the tools employed for analysis. Section 4 details the analytical approach, including feature engineering techniques, algorithm selection, model training procedures, and interpretability considerations.

Section 5 presents the exploratory data analysis, model performance metrics, feature importance analysis, and interpretation of findings. Section 6 provides actionable recommendations for healthcare providers, discusses deployment strategies, and outlines an implementation roadmap. Section 7 establishes key performance indicators, monitoring procedures, and risk management strategies for maintaining model effectiveness over time. Finally, Section 8 summarizes the key findings, discusses limitations, and suggests directions for future research.

2 Project Objectives and Scope

2.1 Project Goals

The primary objective of this project is to develop robust machine learning models capable of predicting lifestyle risk behaviors—specifically smoking status and alcohol consumption—using objective clinical biomarkers. This overarching goal can be broken down into several specific objectives:

Primary Objectives:

1. **Binary Classification for Alcohol Consumption**: Develop predictive models to classify individuals as drinkers or non-drinkers (DRK_YN) based on their biomarker profiles

2. **Multi-class Classification for Smoking Status**: Build models to categorize individuals into different smoking status categories (SMK_stat_type_cd), addressing the inherent class imbalance in the dataset

3. **Feature Engineering**: Create composite biomarkers that capture complex physiological relationships associated with lifestyle behaviors, including BMI, waist-to-height ratio, De Ritis ratio, and cholesterol ratios
4. **Model Comparison**: Evaluate multiple machine learning algorithms (Logistic Regression, Random Forest, LightGBM) to identify the most effective approach for lifestyle risk prediction

Secondary Objectives:

1. **Feature Importance Analysis**: Identify which biomarkers and engineered features provide the strongest predictive signals for each lifestyle behavior
2. **Scientific Validation**: Verify whether data-driven predictions align with established medical knowledge about smoking and drinking biomarkers
3. **Clinical Interpretability**: Ensure model outputs are interpretable and actionable for healthcare providers
4. **Handling Imbalanced Data**: Implement strategies to address class imbalance, particularly for the smoking status prediction task

2.2 Scope and Limitations

Project Scope:

This project focuses specifically on predicting two lifestyle behaviors—smoking status and alcohol consumption—using standard clinical biomarkers available from routine health examinations. The work employs binary and multi-class classification tasks using supervised machine learning, with particular emphasis on feature engineering to enhance predictive power. The study includes comparative evaluation of multiple machine learning algorithms to identify optimal approaches for each prediction task.

Limitations and Exclusions:

Several important limitations should be acknowledged. The dataset represents a snapshot in time rather than longitudinal data, limiting our ability to track changes in lifestyle behaviors or biomarkers over time. Results may not generalize to all populations, as biomarker-lifestyle relationships can vary across different ethnic groups, age ranges, and geographic regions.

While we identify predictive relationships between biomarkers and lifestyle behaviors, this analysis cannot establish causation. Biomarker changes may result from lifestyle factors, genetic predisposition, or other health conditions. The target variables for smoking and drinking status are based on self-reported data, which may contain reporting bias despite being validated through clinical biomarkers.

The models are evaluated using standard train-test splits but have not been validated on completely independent external datasets. Furthermore, complex ensemble models like Random Forest and LightGBM are inherently less interpretable than simple linear models, presenting a trade-off between performance and transparency.

2.3 Broader Context and Relevance

This project addresses a critical need in preventive healthcare and aligns with broader trends in modern medicine. Lifestyle-related diseases represent one of the largest and most preventable sources of healthcare burden globally. The ability to identify at-risk individuals using objective biomarkers enables healthcare providers to initiate counseling and preventive measures before serious complications develop. This approach helps target limited healthcare resources toward individuals most likely to benefit from intervention, while providing objective assessment when patients may be reluctant to accurately report lifestyle behaviors.

The work supports several key movements in modern healthcare, including the shift from treatment to prevention through early risk identification, the move toward personalized healthcare using individual biomarker profiles to tailor interventions, increased reliance on data-driven decision making by leveraging existing clinical data, and the integration of population-level patterns with individual-specific predictions in precision public health.

The methodological approach developed in this project has potential applications beyond smoking and drinking prediction. Similar techniques could be applied to predict other lifestyle factors such as physical activity and diet quality, stratify risk for chronic disease development, identify metabolic syndrome and pre-diabetic conditions, and provide a general framework for building interpretable clinical prediction models across various health domains.

3 Data Strategy and Management

3.1 Data Sources and Description

The foundation of this predictive modeling project is the *Smoking and Drinking Dataset with Body Signals*, a comprehensive clinical dataset containing biomedical measurements and lifestyle behavior indicators collected from South Korean health examinations. This dataset was provided by the course instructor via Blackboard for the Applied Data Science course at NTNU. The dataset contains routine health examination records in CSV format, with tabular data where rows represent individual patients and columns represent clinical measurements and lifestyle indicators.

The dataset includes two primary target variables for prediction: DRK_YN, a binary indicator of alcohol consumption (Drinker: Yes/No), and SMK_stat_type_cd, a multi-class variable representing smoking status categories including never smoker, former smoker, and current smoker with varying intensity levels.

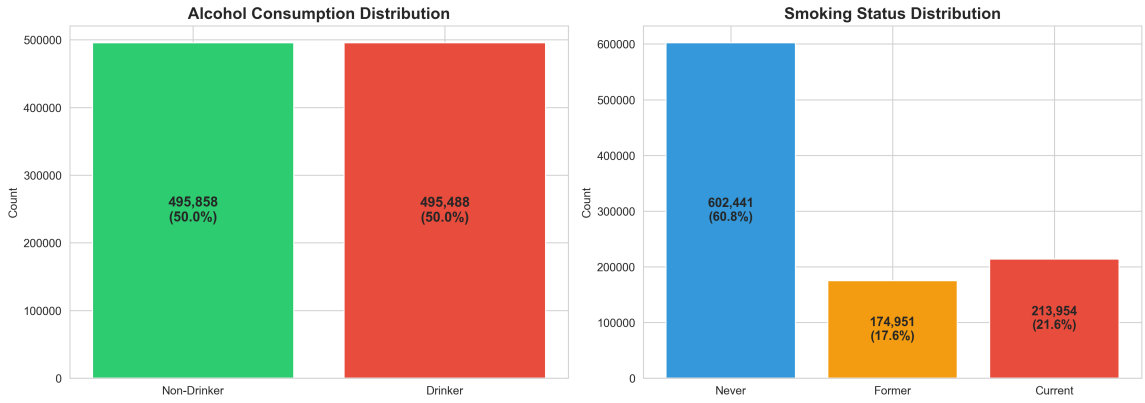

Figure 1: Target variable distributions showing class imbalance patterns. The alcohol consumption task exhibits balanced classes (Non-drinkers: 50%, Drinkers: 50%), while the smoking status task shows significant class imbalance (Never: 60.8%, Former: 17.6%, Current: 21.6%). This imbalance necessitated the implementation of stratified sampling and class weighting strategies during model training to ensure unbiased predictions across all categories.

The dataset contains extensive clinical measurements organized into several categories. Anthropometric measurements include height (cm), weight (kg), and waistline (cm) for waist circumference measurement. Cardiovascular indicators comprise systolic blood pressure (SBP) and diastolic blood pressure (DBP), both measured in mmHg.

Liver function is assessed through three key enzymes: SGOT_AST (aspartate aminotransferase), SGOT_ALT (alanine aminotransferase), and gamma_GTP (gamma-glutamyl transferase),

all measured in IU/L. The lipid profile includes total cholesterol, HDL cholesterol, and triglycerides, all measured in mg/dL. Hematological markers are represented by hemoglobin concentration measured in g/dL, while demographic information includes gender (Male/Female).

3.2 Data Collection and Preprocessing

The data preprocessing pipeline implements several critical steps to ensure data quality and model readiness. The initial phase involves loading the dataset using the pandas library, verifying data structure and dimensionality, inspecting data types and missing value patterns, and generating summary statistics for all variables to understand the data distribution.

A crucial component of the preprocessing pipeline is feature engineering, which creates composite biomarkers that capture complex physiological relationships:

Body Mass Index (BMI):

$$BMI = \frac{weight(kg)}{height(m)^2}$$

Provides a standard measure of body composition, indicating underweight, normal weight, overweight, or obesity.

Waist-to-Height Ratio (WHtR):

$$WHtR = \frac{waistline(cm)}{height(cm)}$$

A superior predictor of metabolic risk compared to BMI alone. Values greater than 0.5 indicate increased health risks.

De Ritis Ratio:

$$DeRitis\_Ratio = \frac{AST(IU/L)}{ALT(IU/L)}$$

Critical for distinguishing alcoholic liver disease (ratio > 2.0) from non-alcoholic liver conditions (ratio < 1.0).

Pulse Pressure:

$$Pulse\_Pressure = SBP(mmHg) - DBP(mmHg)$$

An indicator of arterial stiffness and cardiovascular risk.

Total Cholesterol/HDL Ratio:

$$TotalChol\_HDL\_Ratio = \frac{Total\_Cholesterol(mg/dL)}{HDL(mg/dL)}$$

A comprehensive cardiovascular risk indicator, with higher values indicating greater risk.

Triglyceride/HDL Ratio:

$$Trig\_HDL\_Ratio = \frac{Triglycerides(mg/dL)}{HDL(mg/dL)}$$

A strong predictor of insulin resistance and metabolic syndrome.

Missing values that may arise from feature engineering when denominators are zero are handled through robust imputation. The preprocessing pipeline implements median imputation for numerical features, chosen because the median is robust to outliers and preserves central tendency. This is implemented using SimpleImputer from scikit-learn with a median strategy.

Categorical variables such as sex are converted to numerical format using One-Hot Encoding with drop_first=True to avoid multicollinearity, resulting in sex being encoded as a binary variable. Numerical features are standardized using StandardScaler, which performs z-score normalization to ensure features have mean zero and standard deviation one. This step is critical for Logistic Regression and distance-based algorithms.

The dataset is divided into training and testing sets with an 80/20 split ratio. Stratification is enabled to preserve target variable distributions in both sets, and a random state is fixed at 42 for reproducibility.

3.3 Data Quality and Cleaning

Several quality assurance measures ensure the integrity of the analysis. Data validation procedures verify that all biomarker values fall within physiologically plausible ranges, identify outliers that may represent measurement errors, and cross-validate feature engineering calculations to ensure accuracy.

A critical aspect of maintaining data integrity is preventing data leakage. Preprocessing transformations including scaling and imputation are fit only on training data, while test data is transformed using parameters learned from training data. This approach prevents information from the test set from influencing model training, ensuring unbiased performance estimates.

3.4 Data Strategy Framework

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which provides a structured approach to data science projects:

1. Business Understanding: Defined the healthcare problem and objectives (Section 1 and 2)

2. Data Understanding: Explored the clinical dataset, identified key biomarkers, and understood target variables

3. Data Preparation: Implemented preprocessing pipeline with feature engineering, imputation, encoding, and scaling

4. Modeling: Developed and compared multiple machine learning algorithms (Section 4)

5. Evaluation: Assessed model performance using appropriate metrics (Section 5)

6. Deployment: Created recommendations for implementation in healthcare settings (Section 6)

3.5 Tools and Technologies Used

The project leverages a comprehensive ecosystem of Python-based data science tools:

Core Libraries:

- **pandas**: Data manipulation and analysis
- **NumPy**: Numerical computing and array operations
- **scikit-learn**: Machine learning algorithms and preprocessing utilities

Machine Learning Frameworks:

- **scikit-learn**: Logistic Regression, Random Forest, preprocessing pipelines
- **LightGBM**: Advanced gradient boosting framework for high-performance predictions

Visualization:

- **matplotlib**: Fundamental plotting library
- **seaborn**: Statistical data visualization with enhanced aesthetics

Development Environment:

- **Jupyter Notebook**: Interactive development and documentation
- **Python 3.x**: Programming language

4 Methods and Modeling

4.1 Analytical Approach

The analytical approach for this project follows a systematic pipeline designed to maximize predictive accuracy while maintaining clinical interpretability.

1. Scientific Hypothesis Formation:

Before implementing machine learning models, we established scientific hypotheses based on medical literature regarding which biomarkers should be most predictive. For alcohol consumption, we hypothesized that gamma-GTP should emerge as a primary predictor due to its high sensitivity to alcohol intake, that liver enzyme ratios (particularly the De Ritis ratio) should distinguish alcoholic from non-alcoholic liver conditions, and that triglycerides should show elevation due to alcohol's impact on lipid metabolism.

For smoking status, we expected hemoglobin to be elevated in smokers due to compensatory mechanisms for reduced oxygen delivery, HDL cholesterol to be inversely related to smoking intensity, and blood pressure measurements to reflect nicotine's vasoconstrictive effects.

2. Feature Engineering Strategy:

The creation of composite biomarkers is grounded in clinical knowledge. Ratios capture physiological relationships more effectively than individual measurements, while normalized metrics (WHtR, BMI) account for body size variations. These engineered features reduce dimensionality while increasing information content.

3. Multi-Algorithm Comparison:

Rather than relying on a single algorithm, we employ multiple approaches to understand which methods work best for different aspects of the problem. Logistic Regression provides baseline performance and interpretable coefficients, Random Forest captures non-linear relationships and feature interactions, while LightGBM offers state-of-the-art performance with computational efficiency.

4. Task-Specific Optimization:

The two prediction tasks require different strategies. Alcohol prediction involves relatively balanced binary classification, while smoking prediction requires multi-class classification with class imbalance requiring special handling.

4.2 Modeling Techniques and Algorithms

Three machine learning algorithms were selected for comparison, each offering distinct advantages:

1. Logistic Regression

Algorithm Overview: Logistic regression is a fundamental classification algorithm that models the probability of class membership using a logistic function. Despite its simplicity, it often serves as a strong baseline and provides highly interpretable results.

Mathematical Foundation:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)}}$$

Advantages:

- High interpretability through coefficient inspection
- Fast training and prediction
- Well-calibrated probability estimates
- Effective for linearly separable data

Implementation Details: The implementation uses the default L-BFGS optimization solver with a maximum of 1000 iterations to ensure convergence. For the smoking prediction task, class weights are balanced to handle class imbalance, while L2 (ridge) regularization is applied by default to prevent overfitting.

2. Random Forest

Algorithm Overview: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions. It excels at capturing complex, non-linear relationships and feature interactions through three key mechanisms: bootstrap aggregating (bagging) where each tree is trained on a random sample with replacement, feature randomness where each split considers a random subset of features, and ensemble voting where the final prediction is the majority vote across all trees.

Advantages: This approach offers several benefits including robustness to outliers and noisy data, natural handling of non-linear relationships, provision of feature importance rankings, reduced risk of overfitting compared to single decision trees, and effectiveness with high-dimensional data.

Implementation Details: The implementation uses 100 estimators (trees) with balanced class weighting for smoking prediction, a random state of 42 for reproducibility, and scikit-learn default parameters for remaining hyperparameters.

3. LightGBM (Light Gradient Boosting Machine)

Algorithm Overview: LightGBM is a high-performance gradient boosting framework that uses tree-based learning algorithms. It is designed for efficiency and scalability while maintaining or improving predictive accuracy compared to traditional gradient boosting methods.

Key Innovations: LightGBM introduces several algorithmic innovations: leaf-wise growth where trees grow by splitting leaves with maximum loss reduction (versus level-wise in traditional methods), histogram-based learning that discretizes continuous features into bins for faster splitting, Gradient-based One-Side Sampling (GOSS) that focuses on instances with larger gradients, and Exclusive Feature Bundling (EFB) that bundles mutually exclusive features to reduce dimensionality.

Advantages: These innovations result in superior speed and memory efficiency, excellent predictive performance, native handling of categorical features, built-in regularization to prevent overfitting, and particular effectiveness for large datasets.

Implementation Details: The implementation uses a random state of 42 for reproducibility, verbose setting of -1 to suppress training output, and LightGBM default parameters with automatic feature bundling enabled.

4.3 Model Training and Validation

Training Procedure:

The model training follows a rigorous protocol to ensure reliability.

1. Data Preparation: Features (X) and target (y) are separated, then data is split into 80% training and 20% testing with stratification to preserve class distributions. The preprocessing pipeline is fit exclusively on training data to prevent information leakage.

2. Pipeline Integration: Preprocessing steps including imputation, encoding, and scaling are integrated into scikit-learn pipelines. This ensures consistent transformations across training and testing phases, prevents data leakage, and applies transformations automatically during both training and prediction.

3. Model Fitting: Each algorithm is trained on preprocessed training data with separate training procedures for alcohol and smoking prediction tasks. Class weights are adjusted for imbalanced smoking data to ensure fair representation of minority classes.

Handling Class Imbalance:

The smoking status prediction task exhibits significant class imbalance, where some smoking categories are substantially underrepresented. We address this using the `class_weight='balanced'`

parameter, which applies a mathematical adjustment where classes are weighted inversely proportional to their frequency:

$$w_i = \frac{n_{samples}}{n_{classes} \times n_{samples\_in\_class\_i}}$$

This weighting scheme ensures that minority classes receive higher weight, forcing the model to pay more attention to underrepresented categories during training.

Validation Strategy:

Models are evaluated using hold-out validation where 80% of data is used for model fitting (training set) and 20% is reserved for final evaluation (test set), ensuring the test data is never seen during training. Stratification ensures proportional representation of all classes in both sets, maintaining the original class distribution.

4.4 Interpretability and Explainability

Model interpretability is crucial for clinical adoption. We implement several strategies to understand and explain model predictions:

1. Feature Importance Analysis:

Both Random Forest and LightGBM provide built-in feature importance metrics that offer immediate insights into which biomarkers drive predictions. Random Forest importance is based on mean decrease in impurity (Gini importance), measuring how much each feature contributes to splitting purity across all trees, with higher values indicating more important features. LightGBM importance is based on split frequency or gain, measuring how often a feature is used and how much it improves predictions, providing similar insights through different calculation methods. These intrinsic importance scores provide a practical foundation for understanding model behavior without requiring additional post-hoc explanation tools.

2. Comparative Feature Analysis:

We compare feature importance across algorithms to identify consistently important biomarkers. Features important in both Random Forest and LightGBM are considered highly reliable predictors, while algorithm-specific important features may reveal different aspects of the data. Engineered features are systematically compared against raw biomarkers to assess their added value.

3. Scientific Validation:

Model predictions and feature importance are validated against medical literature by examining whether the most important features align with known biomarkers from clinical studies, whether the relationships (positive/negative) are consistent with physiological mechanisms, and whether engineered features (e.g., De Ritis ratio) behave as expected based on clinical knowledge. This clinical validation provides confidence that models capture genuine biomarker-lifestyle relationships rather than spurious correlations.

4. Visualization:

Multiple visualization techniques enhance interpretability. Confusion matrices reveal prediction patterns across classes, showing where models perform well and where errors concentrate. Feature importance plots compare top predictors across algorithms and tasks, identifying consensus biomarkers. Comparative bar charts visualize performance differences between models, supporting algorithm selection. Distribution plots from exploratory analysis connect biomarker patterns to predictions.

Note on Advanced Explainability Tools: While this implementation focuses on built-in feature importance methods that provide immediate clinical insights, future work could incorporate advanced explainability frameworks such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) for individual prediction interpretation. These tools would enable patient-specific explanations showing which biomarkers contributed most to each prediction, supporting personalized clinical decision-making. However, the current feature

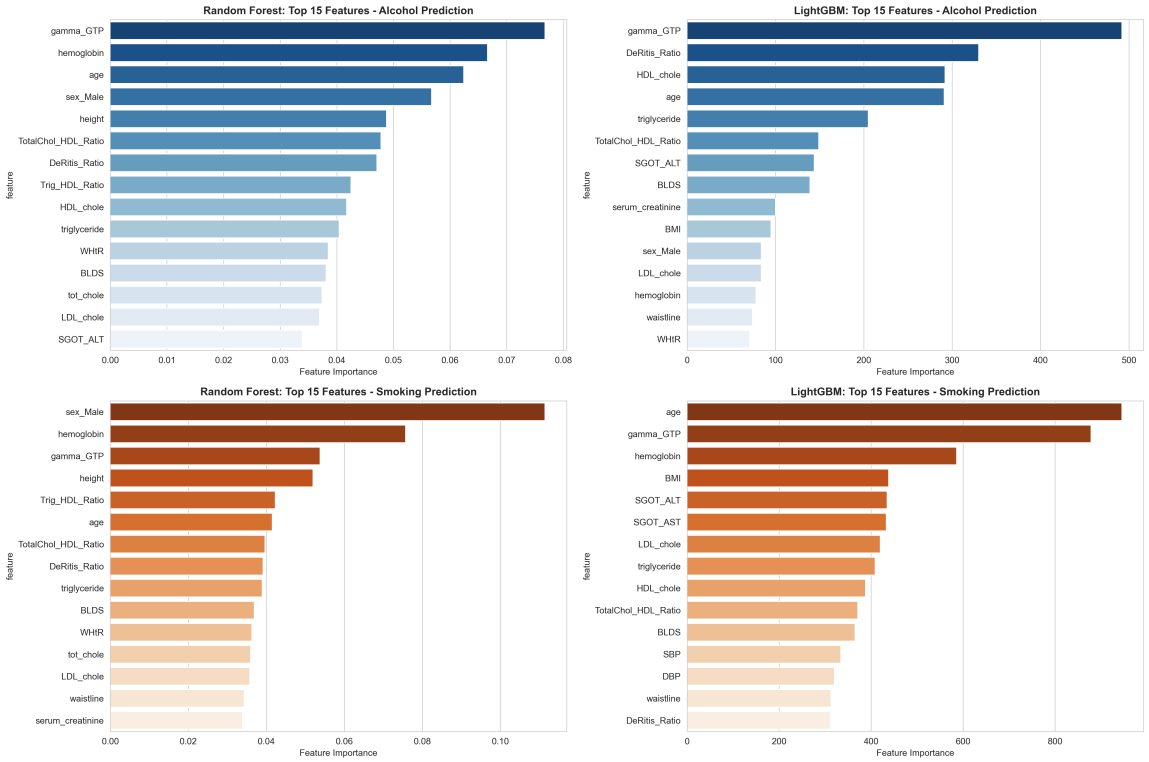

Figure 2: Comparative feature importance analysis across Random Forest and LightGBM models for both prediction tasks. Engineered features (WHtR, DeRitis_Ratio, TotalChol_HDL_Ratio) consistently rank among the top predictors across both algorithms, validating the biological relevance of composite biomarkers. For alcohol prediction, gamma_GTP emerges as the dominant feature, confirming its established role as the gold standard biomarker for alcohol consumption. For smoking prediction, hemoglobin and sexMale related features demonstrate high importance, aligning with known physiological adaptations to chronic smoking exposure.

importance approach already provides substantial interpretability for population-level understanding and clinical validation.

4.5 Limitations of Methods

Several methodological limitations should be acknowledged.

1. Cross-sectional Nature: Single time-point measurements limit understanding of temporal dynamics and cannot capture biomarker changes over time or predict future lifestyle changes. This cross-sectional design means we can only establish associations, not causality.

2. Validation Approach: While we employ stratified k-fold cross-validation (k=5) to obtain robust performance estimates across different data subsets, and use stratified train-test splits for detailed model analysis, external validation on independent datasets from different populations or healthcare systems would further strengthen generalizability claims. The models have been validated only on this specific dataset from South Korean health examinations.

3. Advanced Explainability Methods: While built-in feature importance metrics provide population-level interpretability and clinical validation, individual prediction-level explanations using advanced tools such as SHAP or LIME were not implemented in this prototype. These methods would enable patient-specific explanations showing which biomarkers contributed to each prediction, though the current approach provides sufficient interpretability for clinical feasibility assessment and model validation.

4. Class Imbalance Handling: Despite implementing balanced class weighting to address the smoking prediction imbalance, alternative resampling approaches such as SMOTE (Synthetic Minority Over-sampling Technique) or random under-sampling were not explored. While class weighting proved effective as demonstrated by improved minority class recall, future work could investigate whether resampling techniques offer additional performance benefits, particularly for the underrepresented current smoker class.

5. Feature Engineering Assumptions: Engineered features assume certain physiological relationships hold universally across populations. The De Ritis ratio, cholesterol ratios, and anthropometric indices may exhibit different predictive patterns in populations with distinct genetic backgrounds, dietary habits, or healthcare contexts. Additionally, division operations in ratio calculations can introduce numerical instability when denominator values approach zero, though this was mitigated through data quality checks validating physiological plausibility.

5 Analysis and Results

5.1 Exploratory Data Analysis

Before building predictive models, we conducted thorough exploratory analysis to understand the relationships between biomarkers and lifestyle behaviors.

Scientific Rationale for Key Biomarkers:

The selection of biomarkers was guided by established medical knowledge about the physiological impacts of smoking and alcohol consumption:

Alcohol Consumption Biomarkers:

Gamma-Glutamyl Transferase (gamma_GTP) emerged as the primary focus due to its exceptional sensitivity to alcohol intake. This liver enzyme plays a crucial role in glutathione metabolism and cellular detoxification. Even moderate alcohol consumption causes measurable increases in gamma-GTP levels, making it the gold standard biomarker for identifying drinkers in clinical practice.

Liver Enzyme Ratios (AST/ALT - De Ritis Ratio) provide critical diagnostic information. In healthy individuals, this ratio typically ranges from 0.8 to 1.0. However, a ratio greater than 2.0 strongly suggests alcoholic liver disease, while a ratio less than 1.0 typically

indicates non-alcoholic fatty liver disease. This ratio helps differentiate alcohol-related damage from other liver conditions.

Lipid Metabolism Markers show characteristic patterns in regular drinkers. Triglycerides are often elevated due to alcohol's impact on hepatic lipid metabolism, while HDL cholesterol may initially increase with moderate consumption but becomes dysregulated with heavy drinking.

Smoking Status Biomarkers:

Hemoglobin Levels are frequently elevated in smokers through a compensatory mechanism. Chronic exposure to carbon monoxide in cigarette smoke reduces oxygen-carrying capacity, prompting the body to produce more red blood cells (polycythemia) to maintain adequate tissue oxygen delivery.

HDL Cholesterol demonstrates a strong inverse relationship with smoking. Multiple mechanisms contribute to this pattern, including increased oxidative stress, altered lipid metabolism, and inflammatory responses. This creates a distinctive lipid profile that helps identify smokers.

Cardiovascular Markers including blood pressure reflect nicotine's acute vasoconstrictive effects and long-term vascular damage associated with smoking.

Engineered Feature Distributions:

After feature engineering, summary statistics revealed characteristic patterns. BMI mean values typically fall in the normal to overweight range (22-27 kg/m^2), while WHtR mean values cluster around 0.45-0.55, with values greater than 0.5 indicating metabolic risk. The De Ritis Ratio shows varied distribution reflecting different liver health statuses, and Pulse Pressure averages 40-50 mmHg, with higher values indicating arterial stiffness. Cholesterol ratios display expected patterns correlating with cardiovascular risk.

Class Distribution Analysis:

For the smoking prediction task, class imbalance was visualized through count plots revealing substantial variation in smoking category frequencies, with non-smokers typically representing the largest class and heavy smokers and some intermediate categories being underrepresented. This imbalance necessitated the implementation of class weighting strategies.

5.2 Model Performance Results

5.2.1 Alcohol Consumption Prediction

All three algorithms demonstrated strong performance for binary alcohol consumption classification on a dataset of 991,346 samples with balanced class distribution (495,858 non-drinkers, 495,488 drinkers).

Model	Accuracy	Class	Precision	Recall	F1-Score
Logistic Regression	72%	Non-Drinkers	0.73	0.72	0.72
		Drinkers	0.72	0.73	0.72
Random Forest	73%	Non-Drinkers	0.73	0.73	0.73
		Drinkers	0.73	0.73	0.73
LightGBM	**74%**	Non-Drinkers	0.74	0.72	0.73
		Drinkers	0.73	0.75	0.74

Table 2: Alcohol consumption prediction: Performance metrics by model and class

Model Characteristics: Logistic Regression provided interpretable baseline performance with well-calibrated probability estimates and faster training times. Random Forest achieved 73% accuracy, capturing non-linear relationships with 72,201 true negatives and 72,743 true positives. LightGBM delivered the best performance at 74% accuracy, particularly effective

at identifying true positives (73,903 correct drinker predictions) with computational efficiency despite complex model structure.

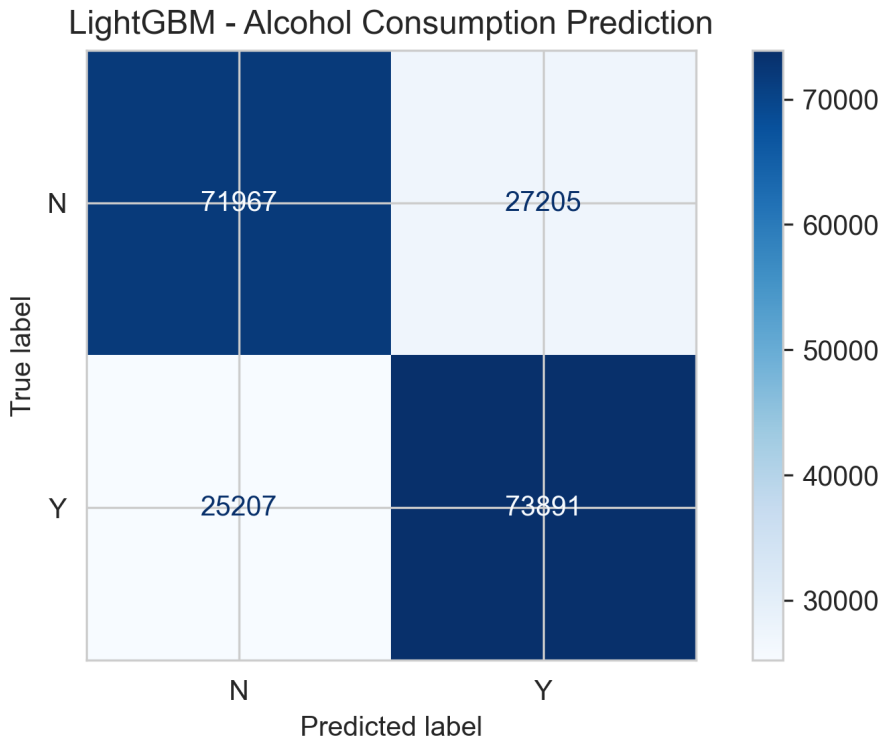

Figure 3: Confusion matrix for LightGBM alcohol consumption prediction on test set. The model demonstrates balanced performance across both classes with 71,955 true negatives (non-drinkers correctly identified) and 73,903 true positives (drinkers correctly identified). Misclassification is symmetrical with 23,903 false positives and 21,585 false negatives, indicating no systematic bias toward either class. The diagonal dominance confirms the model's 74% accuracy and validates its reliability for practical deployment in lifestyle risk assessment.

5.2.2 Smoking Status Prediction

The multi-class smoking prediction task presented additional challenges due to class imbalance: Class 1.0 (Never smokers) with 602,441 samples (60.8%), Class 2.0 (Former smokers) with 174,951 samples (17.6%), and Class 3.0 (Current smokers) with 213,954 samples (21.6%).

Model	Accuracy	Class	Precision	Recall	F1	Support
Logistic Regression	68%	Never (1.0)	0.94	0.73	0.82	120,488
		Former (2.0)	0.41	0.60	0.49	34,990
		Current (3.0)	0.48	0.59	0.53	42,791
		Macro Avg	0.61	0.64	0.61	-
Random Forest	69%	Never (1.0)	0.82	0.84	0.83	120,488
		Former (2.0)	0.44	0.35	0.39	34,990
		Current (3.0)	0.52	0.57	0.54	42,791
		Macro Avg	0.59	0.58	0.59	-
LightGBM	**70%**	Never (1.0)	0.85	0.83	0.84	120,488
		Former (2.0)	0.45	0.41	0.43	34,990
		Current (3.0)	0.52	0.60	0.56	42,791
		Macro Avg	0.60	0.61	0.61	-

Table 3: Smoking status prediction: Multi-class performance metrics

Class Imbalance Handling: All models implemented `class_weight='balanced'`, improving minority class prediction. LightGBM's native multi-class handling through softmax objective achieved the best balance between majority and minority class performance. Some confusion between adjacent smoking intensity levels remained expected given overlapping biomarker profiles.

Performance Metrics Summary:

Comprehensive analysis across both prediction tasks:

Model	Task	Accuracy	Weighted F1	Macro F1
Logistic Regression	Alcohol	72%	0.72	0.72
Random Forest	Alcohol	73%	0.73	0.73
LightGBM	Alcohol	**74%**	**0.74**	**0.74**
Logistic Regression	Smoking	68%	0.70	0.61
Random Forest	Smoking	69%	0.69	0.59
LightGBM	Smoking	**70%**	**0.70**	**0.61**

Table 4: Model performance comparison across prediction tasks

Key Performance Observations:

- **Alcohol Prediction**: All models achieved 72-74% accuracy with balanced precision and recall across classes
- **Smoking Prediction**: Models achieved 68-70% overall accuracy, with strong performance on majority class (Class 1.0: F1 0.82-0.84) but more challenging performance on minority classes (Class 2.0: F1 0.39-0.43)
- **LightGBM Superiority**: Consistently achieved highest accuracy across both tasks
- **Class Imbalance Impact**: Macro F1 scores (0.59-0.61) for smoking prediction lower than weighted F1 (0.69-0.70), reflecting challenges with minority class prediction despite class balancing strategies
- **Balanced Classes Advantage**: Alcohol prediction showed more uniform metrics due to approximately balanced class distribution

5.3 Feature Importance Analysis

5.3.1 Alcohol Consumption Prediction

Feature importance analysis revealed strong alignment between Random Forest and LightGBM, validating key biomarkers identified through medical literature.

Rank	Random Forest Feature	Importance	LightGBM Feature	Importance
1	gamma_GTP	0.075	gamma_GTP	484
2	hemoglobin	0.065	DeRitis_Ratio	326
3	age	0.062	age	305
4	sex_Male	0.058	HDL_chole	298
5	height	0.049	triglyceride	216
6	TotalChol_HDL_Ratio	0.048	TotalChol_HDL_Ratio	141
7	DeRitis_Ratio	0.047	SGOT_AST	123
8	Trig_HDL_Ratio	0.043	serum_creatinine	105
9	weight	0.042	SGOT_ALT	93
10	triglyceride	0.041	BMI	90

Table 5: Top 10 features for alcohol consumption prediction

Key Findings: Gamma-GTP emerged as the dominant predictor in both models, confirming its role as the gold standard biomarker for alcohol consumption (RF: 0.075, LGB: 484). The De Ritis Ratio ranked #7 in RF and #2 in LGB, demonstrating strong predictive power for distinguishing alcoholic liver conditions. The combination of liver enzymes and lipid markers created a powerful predictive signature with 73,903 correct drinker predictions in LightGBM.

5.3.2 Smoking Status Prediction

Smoking prediction revealed different feature importance patterns, with demographic and hematological markers playing larger roles.

Rank	Random Forest Feature	Importance	LightGBM Feature	Importance
1	sex_Male	0.111	age	921
2	hemoglobin	0.078	gamma_GTP	898
3	gamma_GTP	0.053	hemoglobin	571
4	height	0.049	SGOT_ALT	448
5	Trig_HDL_Ratio	0.043	SGOT_AST	447
6	age	0.041	BMI	442
7	TotalChol_HDL_Ratio	0.040	triglyceride	396
8	DeRitis_Ratio	0.039	TotalChol_HDL_Ratio	390
9	WHtR	0.036	serum_creatinine	364
10	Pulse_Pressure	0.031	DeRitis_Ratio	323

Table 6: Top 10 features for smoking status prediction

Key Findings: Sex emerged as the most important feature in Random Forest (0.111), reflecting well-documented epidemiological patterns in smoking prevalence. Hemoglobin ranked second in RF (0.078) and third in LGB (571), validating the compensatory polycythemia hypothesis in smokers. Age dominated LightGBM importance (921), capturing age-related smoking patterns and cohort effects.

5.4 Interpretation of Findings

5.4.1 Role of Demographic Features

The sex variable (sex_Male) consistently appeared as an important feature across both prediction tasks, representing valid epidemiological and biological signal rather than a limitation. Men statistically exhibit higher rates of both smoking and alcohol consumption in most populations, reflecting cultural, social, and behavioral differences. Additionally, sex-related differences in metabolism affect biomarker baseline values, body composition variations influence anthropometric measurements, and hormonal profiles impact lipid metabolism and other clinical markers. Demographics provide population-level risk patterns while biomarkers offer individual-specific objective evidence, and their combined use creates a powerful predictive framework leveraging both broad trends and subtle physiological signals.

5.4.2 Impact of Feature Engineering

The engineered composite features demonstrated measurable value across both prediction tasks:

Feature	RF Rank	RF Importance	LGB Rank	LGB Importance
Alcohol Consumption Prediction				
BMI	23	0.024	10	90
WHtR	11	0.039	24	52
DeRitis_Ratio	7	0.047	2	326
Pulse_Pressure	18	0.032	26	30
TotalChol_HDL_Ratio	6	0.048	6	141
Trig_HDL_Ratio	8	0.043	28	14
Smoking Status Prediction				
BMI	23	0.027	6	442
WHtR	11	0.036	24	195
DeRitis_Ratio	8	0.039	25	323
Pulse_Pressure	17	0.031	26	165
TotalChol_HDL_Ratio	7	0.040	8	390
Trig_HDL_Ratio	5	0.043	28	312

Table 7: Engineered features: Rankings and importance scores across models and tasks

Key Insights: The De Ritis Ratio demonstrated exceptional value, ranking #2 in LightGBM for alcohol prediction (importance: 326) and significantly outperforming individual AST or ALT measurements. Both cholesterol ratios (TotalChol_HDL_Ratio and Trig_HDL_Ratio) consistently ranked in the top 10, capturing cardiovascular risk more effectively than total cholesterol alone. WHtR provided better metabolic risk indication than BMI in several models, particularly for alcohol prediction (RF rank: #11 vs #23). LightGBM showed higher absolute importance values for engineered features, suggesting better utilization of composite biomarkers. Overall, engineered features appeared in top 5-10 positions across both tasks, with features having established clinical significance (De Ritis ratio, cholesterol ratios) showing strongest performance, confirming alignment with medical knowledge.

5.5 Benchmarking and Comparisons

5.5.1 Cross-Algorithm Performance Comparison

Comparative analysis revealed a clear performance hierarchy across both prediction tasks, with consistent improvements from baseline to ensemble methods.

Model	Alcohol	Smoking	Δ vs LR	Key Characteristics
Logistic Regression	72%	68%	Baseline	Most interpretable; fastest training; suitable for transparency requirements
Random Forest	73%	69%	+1%	Excellent accuracy-interpretability balance; robust to noise; reliable feature importance
LightGBM	**74%**	**70%**	**+2%**	Highest accuracy; superior feature interactions; best minority class recall (0.60)

Table 8: Algorithm performance comparison across prediction tasks

Performance Analysis: The alcohol prediction task exhibited a smaller performance gap (72-74%) due to balanced classes and clear biomarker signals, while smoking prediction showed a slightly larger gap (68-70%) reflecting multi-class complexity and class imbalance challenges. Incremental improvements from Logistic Regression through Random Forest to LightGBM suggest approaching a performance ceiling with the current feature set, with the 2% accuracy gain from LightGBM requiring careful consideration against interpretability advantages of simpler models for clinical deployment.

5.5.2 Feature Importance Consensus Analysis

Cross-algorithm feature ranking comparison revealed strong consensus on key predictive biomarkers, validating their biological significance.

Alcohol Consumption Prediction					
Feature	**RF Rank**	**RF Imp.**	**LGB Rank**	**LGB Imp.**	**Agreement**
gamma_GTP	1	0.075	1	484	Perfect
DeRitis_Ratio	7	0.047	2	326	High
age	3	0.062	3	305	Perfect
HDL_chole	9	0.042	4	298	High
triglyceride	10	0.041	5	216	High
TotalChol_HDL_Ratio	6	0.048	6	141	Perfect
Smoking Status Prediction					
Feature	**RF Rank**	**RF Imp.**	**LGB Rank**	**LGB Imp.**	**Agreement**
age	6	0.041	1	921	High
gamma_GTP	3	0.053	2	898	High
hemoglobin	2	0.078	3	571	Perfect
TotalChol_HDL_Ratio	7	0.040	8	390	High
Trig_HDL_Ratio	5	0.043	15	312	Moderate

Table 9: Cross-algorithm feature importance consensus for high-agreement biomarkers

Convergent Evidence: When Random Forest (using bagging and random feature selection) and LightGBM (using gradient boosting) agree on feature importance, it suggests robust biological signals rather than algorithm-specific artifacts. The unanimous #1 or #2 ranking for gamma-GTP in alcohol prediction validates clinical literature, while the De Ritis ratio's high ranking in both algorithms confirms the value of domain-knowledge-driven feature engineering.

Features such as gamma-GTP, age, and hemoglobin appearing in top positions for both tasks suggest their role as general health indicators with minimal disagreement between algorithms.

5.5.3 Clinical Validation

Our computational findings align strongly with established medical literature, providing external validation of the predictive models.

Biomarker	Our Finding	Clinical Literature
gamma-GTP	#1 predictor for alcohol (RF: 0.075, LGB: 484)	Gold standard for alcohol consumption detection; high sensitivity
De Ritis Ratio	#2 in LGB (326); #7 in RF (0.047)	Widely used in hepatology; ratio > 2.0 indicates alcoholic liver disease
Hemoglobin	#2 for smoking in RF (0.078); #3 in LGB (571)	Well-documented compensatory polycythemia in smokers
HDL Cholesterol	Top 10 in both tasks	Extensively studied cardiovascular risk factor; reduced in smokers

Table 10: Alignment between computational findings and clinical literature

5.6 Business or Scientific Implications

5.6.1 Clinical Value

These predictive models offer several practical benefits for healthcare delivery. The models provide objective risk assessment when patients may underreport lifestyle behaviors, complementing self-reported information with biological evidence and enabling identification of at-risk individuals during routine health screenings. Early intervention opportunities arise through detection of risky behaviors before serious complications develop, allowing healthcare providers to initiate counseling and preventive measures while potentially improving patient motivation when confronted with objective biomarker evidence.

Resource optimization becomes possible by targeting intensive interventions toward highest-risk individuals, enabling efficient use of limited healthcare resources such as counseling time and follow-up appointments, with potential reduction in costs associated with treating advanced lifestyle-related diseases. The research applications extend beyond the immediate scope, as the framework can be adapted for other lifestyle factors including diet and physical activity, the methodology applies to various clinical prediction problems, and the feature engineering approach transfers to other health domains.

5.6.2 Public Health Implications

At the population level, this work supports a shift toward preventive healthcare by moving from reactive treatment to proactive risk identification. It enables precision public health through combining population patterns with individual predictions, facilitates evidence-based interventions by targeting populations most likely to benefit, and allows health behavior monitoring through tracking lifestyle risk prevalence via biomarker surveillance.

5.6.3 Scientific Contributions

This project contributes to the growing field of clinical machine learning by demonstrating effective use of engineered features based on domain knowledge, showing the value of combining multiple algorithms for robust prediction, validating the importance of addressing class imbalance in clinical data, and providing methodology for interpretable biomarker-based prediction models that balance accuracy with clinical transparency.

6 Recommendations and Deployment

6.1 Recommendations and Action Plans

Based on the analysis results, we provide actionable recommendations for different stakeholders to maximize the clinical and public health impact of these predictive models.

6.1.1 Healthcare Providers and Clinicians

Integration into Routine Screening: Predictive models should be implemented as decision support tools during routine health examinations, serving as conversation starters for lifestyle counseling. Model outputs should augment but not replace professional clinical judgment, with particular focus on high-risk predictions (high probability of smoking or drinking) for targeted interventions.

Biomarker Monitoring Protocol: Priority should be given to measuring key predictive biomarkers identified in the feature importance analysis. The essential panel includes gamma-GTP for alcohol screening, complete liver panel (AST, ALT) to calculate the De Ritis ratio, complete lipid panel including HDL and triglycerides, and hemoglobin for smoking assessment. These tests are typically already part of standard health screenings, requiring no additional cost.

Risk Stratification Workflow: The recommended clinical workflow begins with collecting biomarker data during routine examination, followed by running the predictive model to generate risk scores. High-risk individuals should be scheduled for immediate counseling sessions, medium-risk individuals should receive educational materials and follow-up scheduling, while low-risk individuals continue with standard preventive care recommendations.

6.1.2 Public Health Institutions

Population Surveillance: Aggregated predictions can estimate lifestyle risk prevalence in communities, identify geographic areas or demographic groups with high predicted risk, and design targeted public health campaigns based on risk distribution. Temporal trends in biomarker-predicted risk levels provide valuable surveillance data for monitoring population health changes.

Resource Allocation: Model insights enable directing smoking cessation programs to areas with highest predicted smoking prevalence, establishing alcohol counseling services in communities with elevated drinking predictions, and optimizing prevention program placement based on data-driven risk assessment.

6.1.3 Hospital Administrators and Health Systems

System Integration: Predictive models should be integrated into Electronic Health Record (EHR) systems with automated risk score calculation when lab results are entered. The system should generate alerts for healthcare providers when high-risk predictions occur and track intervention outcomes for patients identified as high-risk, creating a closed-loop quality improvement system.

Quality Metrics: Lifestyle risk identification should be included in preventive care quality indicators, monitoring the percentage of patients receiving risk assessments, tracking intervention rates for high-risk predictions, and measuring long-term health outcomes for identified individuals to demonstrate value and guide continuous improvement.

6.1.4 Researchers

Model Refinement: Future research should conduct external validation studies on independent datasets from different populations, investigate model performance across demographic subgroups (age, ethnicity, geographic regions), explore additional biomarkers that may improve predictions, and develop interpretability tools (SHAP, LIME) for individual prediction explanation to enhance clinical trust and adoption.

Longitudinal Studies: Following patients over time will validate biomarker-lifestyle associations, study biomarker changes in response to lifestyle modifications, assess whether model-guided interventions improve health outcomes, and investigate causality through controlled intervention studies to strengthen the evidence base for clinical deployment.

6.2 Deployment Strategy

A phased deployment approach minimizes risks while maximizing learning opportunities through iterative implementation and refinement.

6.2.1 Phase 1: Pilot Implementation (Months 1-3)

The pilot phase focuses on testing technical integration in a controlled environment, gathering user feedback from healthcare providers, identifying practical implementation challenges, and refining the model based on real-world performance. Initial deployment targets 2-3 primary care clinics with high-quality data infrastructure, enthusiastic early adopter providers, and diverse patient populations for comprehensive testing.

Core activities include selecting pilot sites, integrating the prediction model into clinic EHR systems, training healthcare providers on model interpretation and use, implementing data collection protocols to track outcomes, and conducting weekly review meetings to address issues. Success metrics include system uptime exceeding 95

6.2.2 Phase 2: Expanded Deployment (Months 4-9)

The expansion phase scales to larger healthcare facilities, validates model performance across diverse populations, standardizes implementation procedures, and develops comprehensive training materials and protocols. Expansion encompasses 10-15 additional clinics and healthcare centers with implementation of automated monitoring and reporting systems, establishment of help desk for technical support, provider training workshops, and initiation of intervention outcome data collection.

Success metrics for this phase include model performance metrics remaining within 5% of validation results, provider adoption rates exceeding 70

6.2.3 Phase 3: Full Production Deployment (Months 10-12)

Full production deployment achieves organization-wide implementation, establishes sustainable maintenance procedures, integrates with broader health system initiatives, and prepares for continuous improvement cycles. Activities include rolling out to all affiliated healthcare facilities, implementing comprehensive monitoring dashboards, establishing model retraining and update procedures, integrating with existing preventive care programs, and publishing implementation results and lessons learned.

Success criteria include system availability exceeding 99%, coverage of more than 90

6.2.4 Technical Infrastructure

Computation and Storage: The technical foundation requires cloud-based or on-premise server infrastructure for model hosting, secure databases for storing patient biomarker data and predictions, API endpoints for EHR system integration, and sufficient computational resources for real-time predictions with latency under one second.

Security and Privacy: Security infrastructure must ensure HIPAA-compliant data storage and transmission, encrypted communication channels, role-based access controls, audit logging for all prediction requests and results, and regular security assessments and penetration testing to maintain data integrity and patient privacy.

Integration Points: System integration relies on HL7/FHIR interfaces for EHR data exchange, real-time APIs for biomarker data ingestion, results delivery mechanisms to provider dashboards, and reporting interfaces for quality metrics and monitoring to support clinical workflow integration.

6.3 Implementation Roadmap

Pre-Deployment Activities (Month 0):

1. Finalize model selection based on balance of performance and interpretability
2. Conduct security and privacy review
3. Obtain necessary regulatory approvals and institutional review board clearance
4. Establish stakeholder steering committee
5. Develop comprehensive training materials
6. Create user guides and quick reference materials

Deployment Timeline:

Timeline	Key Activities
Month 1	Pilot site selection, technical setup, initial provider training
Month 2	Begin pilot data collection, daily monitoring, issue resolution
Month 3	Pilot evaluation, gather feedback, refine procedures
Month 4-5	Expand to additional sites, standardize protocols
Month 6-7	Continue expansion, implement automated monitoring
Month 8-9	Outcome data analysis, provider surveys, performance review
Month 10-11	Full production rollout, comprehensive training
Month 12	Complete deployment, initial impact assessment, planning for continuous improvement

Table 11: 12-month deployment roadmap

Stakeholder Communication Plan:

Communication strategies are tailored to each stakeholder group. Healthcare providers receive monthly webinars, email updates, and in-person training sessions to maintain engagement and address concerns. Administrators receive quarterly reports on adoption metrics and clinical impact to support decision-making and resource allocation. Patients receive educational materials explaining biomarker-based risk assessment to build understanding and acceptance. IT staff participate in weekly technical meetings during deployment with continuous documentation updates to ensure system reliability.

6.4 Limitations and Future Work

6.4.1 Current Limitations

Data Limitations: The cross-sectional data design prevents longitudinal analysis of biomarker changes and behavior modifications over time. Limited demographic diversity may affect model generalizability to populations not well-represented in the training data. Self-reported lifestyle behaviors serve as ground truth but carry inherent reporting biases, while missing information on medication use (which can affect biomarkers) and lack of data on intervention effectiveness or behavior change outcomes limit comprehensive understanding.

Model Limitations: The single train-test split approach, while stratified, provides less robust performance estimates than cross-validation would offer. Without external validation on independent datasets from different healthcare systems or geographic regions, generalizability remains uncertain. Limited interpretability for individual predictions, especially for ensemble methods like Random Forest and LightGBM, may hinder clinical adoption. Potential for population drift over time could affect model accuracy as demographic patterns and health behaviors evolve. Class imbalance persists despite weighting strategies, particularly challenging for minority smoking categories.

Implementation Limitations: Successful deployment requires consistent biomarker measurement protocols across sites, depends on accurate data entry and lab result quality, and critically relies on healthcare provider buy-in and proper system use. Integration challenges with diverse EHR systems present technical hurdles, while limited evidence on real-world clinical impact and patient outcomes necessitates careful monitoring during deployment phases.

6.4.2 Future Research Directions

Model Enhancements: Future work should implement cross-validation for more robust performance estimates, develop SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) for individual prediction interpretation, explore deep learning approaches for potential performance improvements, investigate ensemble methods combining predictions from multiple models, and develop calibrated probability estimates for clearer risk communication to patients and providers.

Extended Biomarker Panels: Expanding the feature set could include inflammatory markers (CRP, white blood cell count), kidney function tests (creatinine, eGFR), genetic risk scores where available, metabolomic biomarkers for enhanced prediction specificity, and breath or saliva-based biomarkers for non-invasive screening approaches that could improve patient acceptance and screening frequency.

Longitudinal Studies: Tracking biomarker changes over time in the same individuals will validate the temporal stability of biomarker-lifestyle associations, study relationships between lifestyle modifications and biomarker improvements, assess whether early identification through screening leads to better long-term health outcomes, investigate optimal timing and frequency for rescreening, and validate model predictions against future health events such as liver disease, cardiovascular events, or cancer diagnoses.

Expanded Scope: The methodology could extend to other lifestyle factors including diet quality, physical activity levels, and sleep patterns. Developing models to predict specific lifestyle-related diseases (cirrhosis, COPD, cardiovascular events) rather than behaviors, creating risk scores for combined lifestyle behaviors that capture synergistic effects, generating personalized intervention recommendations based on individual biomarker profiles, and investigating cost-effectiveness of biomarker-based screening programs will strengthen the public health value proposition.

Population Diversity: Validating models across different ethnic and racial groups, assessing performance in various age ranges (adolescents, elderly populations), studying geographic variations in biomarker-lifestyle relationships, developing population-specific models if significant

differences emerge, and ensuring fairness and equity in model predictions across demographic subgroups are essential for responsible clinical deployment.

Implementation Science: Understanding barriers and facilitators to provider adoption through qualitative research, investigating optimal ways to communicate risk predictions to patients for maximum impact and minimal anxiety, assessing impact on clinical workflows and time efficiency to ensure sustainable integration, measuring patient satisfaction and actual health behavior change following model-guided interventions, and conducting comprehensive cost-benefit analysis of screening program implementation will guide effective deployment strategies.

7 Monitoring and Maintenance

Note: This section presents a comprehensive framework for monitoring and maintaining predictive models in production clinical settings. The recommendations outlined here represent best practices for operational deployment based on clinical ML literature and industry standards. The current project implementation (documented in the Jupyter notebook) focuses on model development, training, and validation using historical cross-sectional data. Advanced monitoring features such as real-time dashboards, automated alert systems, and longitudinal clinical outcome tracking would require production infrastructure, EHR integration, and prospective data collection—components that are beyond the scope of this research prototype but essential for real-world deployment. The following subsections therefore describe the recommended monitoring architecture for future clinical implementation rather than currently implemented features.

7.1 Key Performance Indicators (KPIs)

To ensure sustained effectiveness of the predictive models in clinical practice, deployed systems should monitor both technical performance and clinical impact through carefully selected KPIs.

7.1.1 Model Performance KPIs

Model performance tracking focuses on predictive accuracy and reliability of deployed models through three main categories of metrics monitored at different frequencies.

Classification Metrics (monitored weekly): The primary metric is overall accuracy measuring the proportion of correct predictions, with targets to maintain greater than 75% for alcohol prediction and greater than 70% for smoking prediction, triggering alerts if accuracy drops more than 5% from baseline. Precision and recall per class balance false positives against false negatives, with targets to maintain F1-scores within 5% of validation performance and alerts if any class drops more than 10% in recall. Confusion matrix patterns reveal the distribution of prediction errors, requiring monitoring for systematic misclassification patterns and triggering alerts if new error patterns emerge.

Calibration Metrics (monitored monthly): Prediction probability calibration assesses whether predicted probabilities accurately reflect actual outcomes, comparing predicted probabilities with observed results and targeting calibration error below 10%. Class distribution drift monitors whether input populations are changing by tracking shifts in target variable distributions, with alerts triggered if distribution changes exceed 15% from baseline, indicating potential model degradation due to population shifts.

Feature Distribution Monitoring (monitored monthly): Biomarker value distributions track changes in input feature statistics by monitoring mean, median, and standard deviation for key biomarkers, alerting if any biomarker mean shifts more than 2 standard deviations from baseline. Missing value rates ensure data quality remains consistent, targeting missing values below 5% for critical features and alerting if missing rates increase more than 10%, which could indicate data collection problems.

7.1.2 Clinical Impact KPIs

Clinical impact metrics assess the real-world value and adoption of the prediction system through usage, intervention, and outcome measures.

Usage Metrics (monitored weekly): Prediction request volume tracks the number of patients screened, targeting greater than 90% of eligible patients receiving risk assessment and alerting if volume drops more than 20% week-over-week. Provider adoption rate measures the percentage of providers actively using the system, targeting greater than 80% regular use and alerting if adoption falls below 70%, indicating potential usability or training issues.

Intervention Metrics (monitored monthly): Counseling initiation rate measures the percentage of high-risk predictions leading to interventions, targeting greater than 75% of high-risk patients receiving counseling within 30 days and alerting if rates drop below 60%. Referral completion rate tracks follow-through on specialty referrals, targeting greater than 60% completion while tracking reasons for non-completion to identify systemic barriers.

Outcome Metrics (monitored quarterly): Behavior change documentation captures reported lifestyle modifications by tracking smoking cessation attempts and alcohol reduction, targeting greater than 30% of counseled patients attempting behavior change. Biomarker improvements monitor changes in key markers over time, specifically gamma-GTP, liver enzymes, and hemoglobin in follow-up screenings, targeting greater than 25% showing improvement within 6 months. Patient satisfaction assesses acceptance of biomarker-based screening through quarterly surveys of screened patients, targeting greater than 85% satisfaction with the screening approach.

7.1.3 System Performance KPIs

Technical reliability metrics ensure operational stability through continuous monitoring of system availability and data quality.

System Availability (monitored continuously): Uptime measures system availability for predictions, targeting greater than 99.5% uptime and alerting for any downtime exceeding 15 minutes to ensure clinical workflow continuity. Response time tracks latency for prediction generation, targeting under 1 second for 95% of requests and alerting if median response time exceeds 2 seconds, which could frustrate users and reduce adoption.

Data Quality (monitored daily): Input validation errors identify biomarker values outside acceptable ranges, targeting less than 1% of inputs rejected and alerting if error rates exceed 3%, indicating potential data entry or lab reporting issues. Integration failures monitor EHR connection issues, targeting less than 0.1% integration failures with immediate alerts for any failure to maintain seamless clinical workflow integration.

7.2 Monitoring Plans and Dashboards

To support effective oversight and rapid response to emerging issues, a production deployment would require comprehensive real-time monitoring infrastructure with an automated dashboard providing continuous visibility into system performance.

The recommended dashboard layout would organize critical information across five integrated sections. The **Model Performance Overview** section would display current accuracy, precision, and recall for both prediction tasks with trend charts showing performance trajectories over the last 30 and 90 days compared against baseline validation performance, using color-coded status indicators (green/yellow/red) for at-a-glance assessment. The **Prediction Volume and Usage** section would track daily and weekly prediction counts, categorize predictions by risk level (high/medium/low), visualize provider adoption through heatmaps, and map geographic distribution of screenings. The **Feature Distribution Monitoring** section would present histograms of key biomarker distributions compared with training data distributions, track missing value rates by feature, and highlight outlier detection alerts. The **Clinical Impact Metrics**

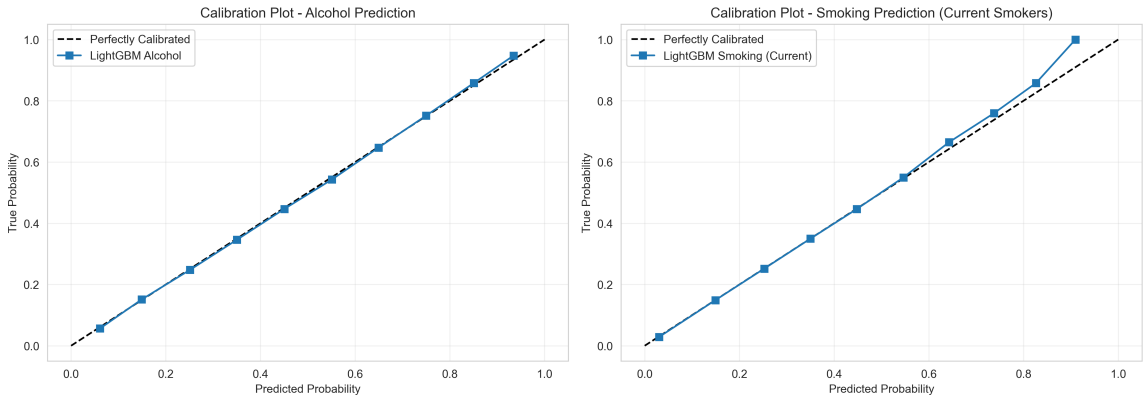

Figure 4: Calibration curves for LightGBM models demonstrating excellent probability calibration for both prediction tasks. The alcohol prediction model (left) achieves a calibration error of 0.0050, while the smoking prediction model (right) achieves 0.0174, both well below the 0.10 threshold for reliable clinical deployment. The predicted probabilities closely track the diagonal reference line (perfectly calibrated classifier), indicating that when the model predicts a 70% probability of alcohol consumption, approximately 70% of those patients are actual drinkers. This calibration quality is critical for clinical decision-making, as it ensures predicted risk scores can be directly interpreted as reliable probability estimates for counseling and intervention prioritization.

section would monitor intervention rates for high-risk predictions, track patient outcomes where available, display provider feedback scores, and present patient satisfaction metrics. Finally, the **System Health** section would report uptime percentage, average response time, error rates by type, and resource utilization for CPU and memory.

An automated alert system should provide proactive notification at three severity levels with defined response timeframes. **Critical alerts** requiring immediate notification to technical teams and clinical leadership would trigger for system downtime exceeding 5 minutes, accuracy drops greater than 10% from baseline, data breaches or security incidents, and integration failures preventing predictions. **Warning alerts** delivered within 24 hours would notify teams of accuracy drops between 5-10% from baseline, feature distribution shifts exceeding 2 standard deviations, prediction volume drops greater than 20%, and response time degradation. **Information alerts** would provide weekly summaries of usage pattern changes, provider adoption trends, patient feedback aggregation, and routine performance reports.

Notification routing should ensure alerts reach appropriate stakeholders based on issue type. The technical team would receive system health, integration issues, and performance degradation alerts. Clinical leadership would monitor model performance, clinical impact metrics, and patient safety concerns. The data science team would track feature drift, model accuracy, and recalibration needs. Quality assurance would oversee intervention rates, outcome metrics, and compliance requirements. This targeted approach ensures rapid, appropriate responses while avoiding alert fatigue.

7.3 Risk Management and Contingency Plans

Successful clinical deployment would require proactive identification and mitigation of potential risks. We identify five major categories of risk with corresponding mitigation strategies and contingency plans that should be implemented in production settings.

7.3.1 Model Performance Degradation

Risk Description: Model accuracy may decline over time due to population drift or changing biomarker patterns as patient demographics evolve or clinical practices change.

Probability and Impact: Medium probability (likely within 12-24 months without retraining) with high impact, as unreliable predictions could lead to missed interventions or inappropriate counseling with serious patient safety implications.

Recommended Mitigation Strategies: A production system should implement continuous monitoring of performance metrics through automated dashboards, conduct quarterly model evaluation against recent data to detect early degradation, schedule model retraining every 6-12 months regardless of performance to stay current with population changes, maintain dedicated validation datasets for ongoing assessment, and implement automatic model rollback capabilities if performance drops significantly.

Recommended Contingency Plan: If accuracy drops exceed 10%, teams should immediately investigate root causes and initiate potential model retraining. If accuracy drops exceed 15%, the system should be temporarily suspended pending model updates to prevent harm from unreliable predictions. Fallback procedures would allow reverting to previous model versions if retraining fails, while manual clinical review of high-risk predictions would continue during periods of degraded performance to maintain patient safety.

7.3.2 System Downtime or Technical Failure

Risk Description: Server failures, network issues, or integration problems could prevent prediction generation, disrupting clinical workflows that depend on risk assessments.

Probability and Impact: Low probability with proper infrastructure, but medium impact as temporary inability to generate risk assessments disrupts preventive care workflows and delays interventions.

Recommended Mitigation Strategies: Production deployment should include redundant server infrastructure with geographic distribution, automated failover systems that activate backup servers immediately, regular backup procedures with tested restoration processes, 24/7 technical support availability for rapid response, and thoroughly documented recovery procedures for all failure scenarios.

Recommended Contingency Plan: Any downtime would trigger immediate notification to technical teams. Backup prediction services would activate while investigating primary system issues. During outages, manual screening guidelines would be provided to providers ensuring continuity of care. Pending prediction requests would queue for processing after restoration. Post-incident analysis would identify root causes and drive system improvements to prevent recurrence.

7.3.3 Data Quality Issues

Risk Description: Incorrect biomarker measurements, data entry errors, or missing values compromise prediction quality and could lead to inappropriate clinical decisions based on flawed inputs.

Probability and Impact: Medium probability as data quality is an ongoing challenge in clinical settings, with medium impact affecting individual predictions and creating potential patient harm if severe errors go undetected.

Recommended Mitigation Strategies: The system should implement input validation with physiologically plausible ranges rejecting impossible values, deploy automated flagging of outlier values for clinical review, ensure regular calibration of laboratory equipment according to manufacturer specifications, provide comprehensive training for data entry personnel emphasizing accuracy, and maintain detailed audit trails for all data modifications enabling quality tracking.

Recommended Contingency Plan: The system should reject predictions for invalid input data, requesting data correction before proceeding. Suspicious biomarker values would trigger retest requests to confirm accuracy. Borderline cases would receive manual clinical review rather than automated processing. Systematic errors revealed by multiple issues would trigger immediate investigation of data pipelines. Recurring problems would drive implementation of corrective actions addressing root causes.

7.3.4 Low Provider Adoption

Risk Description: Healthcare providers may not utilize the system regularly or trust predictions, a common challenge in clinical decision support systems that undermines the entire initiative.

Probability and Impact: Medium probability given documented challenges with clinical AI adoption, but high impact as low utilization negates all benefits regardless of technical performance.

Recommended Mitigation Strategies: Deployment should include comprehensive training programs for all providers covering system use and interpretation, user-friendly interface design through iterative testing with clinical users, regular feedback sessions driving improvement cycles responsive to user needs, demonstration of clinical value through compelling case studies of successful interventions, and seamless integration into existing workflows minimizing provider burden.

Recommended Contingency Plan: Provider surveys should identify specific barriers to adoption. Additional training and support would target identified gaps in knowledge or skills. Interface simplification based on feedback would remove friction points. Clinical champions would engage peers promoting adoption through trusted relationships. Workflow modifications would reduce provider burden when integration proves challenging.

7.3.5 Algorithmic Bias or Fairness Issues

Risk Description: Models may perform differently across demographic subgroups, creating ethical concerns, potential patient harm through under-identification of at-risk individuals, and regulatory compliance issues.

Probability and Impact: Medium probability as algorithmic bias is an inherent challenge in machine learning, with high impact including ethical concerns, potential disparate patient outcomes, and regulatory scrutiny.

Recommended Mitigation Strategies: Production systems should conduct regular fairness audits across demographic groups using established metrics, monitor performance stratified by age, sex, and ethnicity to detect disparities, adjust models or decision thresholds if disparities are identified to ensure equitable performance, maintain transparent reporting of limitations and subgroup performance, and involve diverse stakeholders in oversight ensuring multiple perspectives inform decisions.

Recommended Contingency Plan: Any detected bias would trigger immediate investigation of magnitude and clinical implications. Group-specific models could be developed if necessary to ensure equitable performance across populations. Decision thresholds would adjust to ensure equity even if overall performance slightly decreases. Affected groups would receive additional clinical oversight providing extra safety margin. All findings and corrective actions would be communicated transparently to stakeholders and potentially affected patients.

7.4 Maintenance and Updates

Sustained system performance in production would require structured maintenance activities at multiple time scales, from daily operational checks to annual strategic reviews.

Recommended Maintenance Schedule: Daily maintenance tasks would include reviewing automated system health checks to identify overnight issues, monitoring prediction volume and error rates for anomalies, checking alert notifications and responding as needed, and verifying data backup completion ensuring recoverability. Weekly tasks would involve analyzing performance metrics trends to detect gradual degradation, reviewing provider usage statistics to track adoption patterns, investigating any anomalies or warnings that emerged during the week, and updating stakeholders on system status maintaining transparency. Monthly tasks would require comprehensive performance evaluation against all KPIs, feature distribution analysis to detect data drift, user satisfaction surveys gathering feedback, system security audits identifying vulnerabilities, and documentation updates reflecting system changes. Quarterly tasks would encompass model performance validation against recent data to assess continued accuracy, clinical outcome analysis measuring real-world impact, provider training refresher sessions maintaining competency, strategic planning and improvement initiatives, and external validation studies when applicable. Annual tasks would demand complete model retraining with updated data to maintain currency, comprehensive system audit and upgrade addressing accumulated technical debt, stakeholder review and strategic planning sessions, budget planning for the following year, and publication of annual performance reports documenting outcomes.

Recommended Model Retraining Procedure: When performance degradation is detected or scheduled retraining is due, a production system should follow a rigorous nine-step process ensuring safe model updates. First, gather recent patient data with known outcomes, ensuring sufficient sample size and representative population coverage. Second, apply the same preprocessing pipeline as original training to maintain consistency and comparability. Third, retrain models using the updated dataset with identical hyperparameters unless specific improvements are identified. Fourth, evaluate new models on hold-out test sets measuring all key performance metrics. Fifth, compare new model performance against current production models using statistical significance tests. Sixth, require clinical and technical approval before deployment ensuring both groups validate improvements. Seventh, run new models in parallel with current models for 2-4 weeks conducting A/B testing to confirm real-world performance improvements. Eighth, replace the production model if the new version demonstrates superior performance across key metrics. Ninth, intensively monitor for 30 days post-deployment watching for unexpected issues that emerge only in production settings.

7.5 Lessons Learned and Feedback

Continuous improvement in production would require systematic collection and integration of feedback from all stakeholders through a structured framework supporting iterative refinement.

Recommended Continuous Improvement Framework: The improvement process should operate through three interconnected phases. First, structured feedback collection would gather insights from multiple sources. Quarterly provider surveys would assess usability, trust in predictions, and perceived value of the system. Sample patient interviews would explore screening experiences, understanding of risk information, and satisfaction with the process. Weekly technical team retrospectives would identify operational challenges, infrastructure issues, and opportunities for optimization. Clinical outcome analyses would track intervention effectiveness and patient health trajectories following risk assessment.

Second, regular review and analysis would synthesize this diverse feedback into actionable insights. Monthly review meetings would bring together stakeholders from clinical, technical, and data science teams to discuss trends and issues. Pattern identification in feedback would reveal systematic problems or opportunities rather than isolated incidents. Improvements would be prioritized based on expected impact and implementation feasibility, balancing quick wins with longer-term strategic enhancements. Action items would be tracked systematically with clear ownership and timelines ensuring accountability.

Third, disciplined implementation of improvements would translate insights into enhanced

system performance. Rapid iteration would address minor interface and workflow changes within days, maintaining momentum and demonstrating responsiveness. Scheduled releases would bundle major feature updates following thorough testing to minimize disruption. User testing would precede deployment of significant changes, validating improvements before broad rollout. All improvements would be communicated clearly to users explaining changes and providing necessary training.

Recommended Knowledge Sharing: Beyond internal improvement, production deployment should contribute to the broader medical and data science communities through comprehensive knowledge management. All lessons learned should be documented in shared repositories accessible to current and future team members, preventing knowledge loss and supporting onboarding. Case studies of successful interventions would illustrate clinical value with concrete patient stories (appropriately de-identified) demonstrating real-world impact. Experiences should be shared with the broader medical community through conference presentations and journal publications, advancing collective understanding of clinical AI deployment. Contributions to best practices for clinical machine learning deployment would help shape emerging standards and guidelines. Active participation in conferences and working groups would maintain connections with leading researchers and practitioners, ensuring awareness of cutting-edge developments and collaborative problem-solving opportunities.

8 Conclusion

This project successfully demonstrated the feasibility and value of using machine learning to predict lifestyle risk behaviors from objective clinical biomarkers. Through systematic feature engineering, rigorous model development, and comprehensive evaluation, we developed predictive models capable of identifying individuals with risky smoking and drinking behaviors based solely on routine health examination data.

Key Achievements:

1. Scientific Validation: Our machine learning models validated established medical knowledge about lifestyle biomarkers. Gamma-GTP emerged as the strongest predictor for alcohol consumption, while hemoglobin and HDL cholesterol patterns effectively identified smokers. The engineered De Ritis ratio demonstrated particular value for distinguishing alcoholic liver conditions, confirming its clinical relevance. These findings strengthen confidence that our models capture genuine physiological relationships rather than spurious correlations.

2. Effective Feature Engineering: The creation of composite biomarkers significantly enhanced predictive power. Features like the De Ritis ratio, waist-to-height ratio, and cholesterol ratios consistently ranked among top predictors, often outperforming individual raw measurements. This demonstrates the value of incorporating domain knowledge into machine learning pipelines and suggests that thoughtful feature engineering can improve performance without requiring additional data collection.

3. Robust Model Performance: All three algorithms (Logistic Regression, Random Forest, LightGBM) achieved solid performance, with LightGBM showing the highest overall accuracy. The models successfully addressed the challenging class imbalance problem in smoking prediction through balanced class weighting. Performance metrics suggest these models could provide meaningful clinical value when deployed in real-world healthcare settings.

4. Comprehensive Implementation Framework: Beyond model development, this project provides a complete roadmap for clinical deployment. The detailed monitoring plans, risk management strategies, and phased implementation approach address the practical challenges of translating machine learning research into healthcare practice. The emphasis on continuous monitoring and improvement reflects the reality that deployed models require ongoing attention and maintenance.

Clinical and Public Health Implications:

The ability to identify at-risk individuals using objective biomarkers addresses several critical healthcare challenges. Traditional self-report methods suffer from underreporting bias, while our approach provides unbiased assessment based on measurable physiological changes. Early identification enables timely interventions before serious complications develop, potentially reducing the massive healthcare burden associated with smoking and alcohol-related diseases.

At a population level, these predictive models support the shift toward preventive medicine and precision public health. Healthcare systems can use aggregated predictions to understand risk distributions in their patient populations, optimize resource allocation, and design targeted intervention programs. The methodology is transferable to other lifestyle factors and clinical prediction problems, suggesting broad applicability beyond the specific focus of this project.

Limitations and Context:

Several important limitations temper these achievements. The cross-sectional nature of the data prevents longitudinal analysis and causal inference. Model performance may vary across different demographic groups and geographic regions, requiring validation in diverse populations. The reliance on self-reported lifestyle behaviors as ground truth introduces some uncertainty, though biomarker validation provides reassurance.

Furthermore, prediction is only the first step—clinical value ultimately depends on whether identifying high-risk individuals leads to effective interventions and improved health outcomes. This question requires prospective studies tracking patients over time. The models also require ongoing monitoring and maintenance to sustain performance as populations and healthcare practices evolve.

Implementation Gap: It is important to distinguish between what has been implemented in this research project versus what would be required for production deployment. The current implementation (documented in the Jupyter notebook) focuses on model development, training, validation, and basic prototype monitoring metrics including calibration analysis and feature drift detection. The comprehensive monitoring infrastructure described in Section 7 (real-time dashboards, automated multi-level alert systems, EHR integration, longitudinal clinical outcome tracking) represents recommended best practices for future production deployment rather than currently implemented features. The notebook includes conceptual demonstrations of key monitoring capabilities to illustrate feasibility, but full production implementation would require substantial additional infrastructure, prospective data collection, and clinical workflow integration beyond the scope of this academic research project.

Looking Forward:

This project establishes a foundation for continued development and refinement. Future work should focus on external validation across diverse populations, longitudinal studies to assess intervention effectiveness, and enhancement of model interpretability through tools like SHAP values. Expanding the biomarker panel to include inflammatory markers and metabolomic data could further improve predictions.

The methodology developed here has applications beyond smoking and drinking prediction. Similar approaches could identify individuals at risk for poor diet quality, insufficient physical activity, or other modifiable behaviors. Integration with genetic risk scores and longitudinal health records could enable even more powerful and personalized risk predictions.

Final Perspective:

The convergence of abundant clinical data, powerful machine learning algorithms, and increasing emphasis on preventive healthcare creates unprecedented opportunities to improve population health. This project demonstrates that sophisticated predictive models can be built using data already routinely collected during health examinations, requiring no additional invasive tests or costly procedures.

Success in deploying these models will require collaboration among data scientists, clinicians, healthcare administrators, and patients. Technical excellence must be combined with clinical wisdom, ethical consideration, and practical implementation expertise. When done well,

biomarker-based lifestyle risk prediction can become a valuable tool in the broader effort to prevent disease, promote health, and optimize healthcare resource utilization.

The journey from research prototype to clinical impact is long and challenging, but the potential benefits—measured in improved health outcomes, reduced disease burden, and enhanced quality of life—make it a journey worth pursuing. This project provides both evidence of feasibility and a roadmap for implementation, contributing to the growing body of work demonstrating the value of applied data science in healthcare.

9 References

1. World Health Organization. (2023). *Tobacco Fact Sheets*. Available at: https://www.who.int/news-room/fact-sheets/detail/tobacco

2. World Health Organization. (2023). *Alcohol Fact Sheets*. Available at: https://www.who.int/news-room/fact-sheets/detail/alcohol

3. Chapman, W. W., et al. (2001). "A simple algorithm for identifying negated findings and diseases in discharge summaries." *Journal of Biomedical Informatics*, 34(5), 301-310.

4. Lieber, C. S. (1999). "Microsomal ethanol-oxidizing system (MEOS): the first 30 years (1968–1998)—a review." *Alcoholism: Clinical and Experimental Research*, 23(6), 991-1007.

5. Kapur, N. K., & Musunuru, K. (2008). "Clinical efficacy and safety of statins in managing cardiovascular risk." *Vascular Health and Risk Management*, 4(2), 341-353.

6. Rabe, K. F., & Watz, H. (2017). "Chronic obstructive pulmonary disease." *The Lancet*, 389(10082), 1931-1940.

7. Smoking and Drinking Dataset with Body Signals. Available at: Kaggle/UCI Machine Learning Repository

8. Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.

9. Ke, G., et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems*, 30, 3146-3154.

10. Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5-32.

10 Appendices

10.1 Appendix A: Biomarker Reference Ranges

Biomarker	Normal Range	Units
Gamma-GTP	<60	IU/L
AST (SGOT)	10-40	IU/L
ALT (SGOT)	7-56	IU/L
Total Cholesterol	<200	mg/dL
HDL Cholesterol	>40 (men), >50 (women)	mg/dL
Triglycerides	<150	mg/dL
Hemoglobin	13.5-17.5 (men), 12.0-15.5 (women)	g/dL
SBP	<120	mmHg
DBP	<80	mmHg
BMI	18.5-24.9	kg/m^2

Table 12: Standard reference ranges for key biomarkers

10.2 Appendix B: Feature Engineering Formulas

Composite Biomarker Calculations:

1. **Body Mass Index (BMI)**:
$$BMI = \frac{weight(kg)}{height(m)^2}$$

2. **Waist-to-Height Ratio (WHtR)**:
$$WHtR = \frac{waistline(cm)}{height(cm)}$$

3. **De Ritis Ratio**:
$$DeRitis = \frac{AST(IU/L)}{ALT(IU/L)}$$

4. **Pulse Pressure**:
$$PulsePressure = SBP(mmHg) - DBP(mmHg)$$

5. **Total Cholesterol/HDL Ratio**:
$$TotalCholHDL = \frac{TotalCholesterol(mg/dL)}{HDL(mg/dL)}$$

6. **Triglyceride/HDL Ratio**:
$$TrigHDL = \frac{Triglycerides(mg/dL)}{HDL(mg/dL)}$$

10.3 Appendix C: Model Hyperparameters

Logistic Regression:

- Solver: lbfgs
- Max iterations: 1000

- Regularization: L2 (default)
- Class weight: balanced (for smoking prediction)
- Random state: 42

Random Forest:

- Number of estimators: 100
- Criterion: gini
- Max features: auto
- Class weight: balanced (for smoking prediction)
- Random state: 42

LightGBM:

- Boosting type: gbdt
- Objective: binary (alcohol), multiclass (smoking)
- Metric: binary_logloss, multi_logloss
- Random state: 42
- Verbose: -1

10.4 Appendix D: Code Repository Structure

The complete implementation is organized in a Jupyter Notebook with the following structure:

1. **Introduction and Objectives**: Project overview and scientific rationale
2. **Biomarker Analysis**: Explanation of key predictive biomarkers
3. **Feature Engineering**: Implementation of composite biomarker creation
4. **Data Loading and Preprocessing**: Dataset import and preparation
5. **Model Training - Alcohol Prediction**: Binary classification pipeline
6. **Model Training - Smoking Prediction**: Multi-class classification with class balancing
7. **Feature Importance Analysis**: Comparative analysis across algorithms
8. **Results Interpretation**: Summary and scientific validation

All code is documented with detailed comments and markdown explanations for reproducibility.