Smoker Detection from Bio-Signals Using Confidence-Weighted Ensemble Learning

Chandana Vinay Kumar
Arizona State University
ID: 1237227955
cvinayku@asu.edu

Dileep Pabbathi
Arizona State University
ID: 1236546573
dpabbat1@asu.edu

Ajay Bingi
Arizona State University
ID: 1237907114
abingi1@asu.edu

Aarya Pendharkar
Arizona State University
ID: 1238006590
apendha1@asu.edu

***Abstract*—Smoking status prediction from biomedical signals represents a critical healthcare challenge with significant implications for preventive medicine and personalized treatment planning. This paper presents a novel Confidence-Weighted Ensemble approach that combines gradient boosting methods (XGBoost, LightGBM, CatBoost) with deep learning architectures (TabNet, NODE) using calibration-based weighting rather than traditional accuracy-based or equal-weight averaging. Our key innovation lies in weighting ensemble components by their calibration quality—measured through Expected Calibration Error (ECE) and Brier Score—ensuring that models with more reliable probability estimates contribute proportionally more to final predictions. We engineer 51 domain-specific features from 22 original biomarkers, capturing body composition indices, lipid ratios, liver function indicators, and cardiovascular risk factors. Experimental evaluation on 159,256 health records demonstrates that our Brier-Optimized ensemble achieves a ROC-AUC of 0.8630, outperforming individual models while maintaining superior calibration (ECE: 0.0114). The SHAP-based interpretability analysis reveals that hemoglobin levels, height, and liver enzyme ratios are the strongest predictors. Our approach provides healthcare practitioners with both accurate predictions and reliable confidence estimates, essential for clinical decision-making.**

***Index Terms*—Ensemble Learning, Smoker Detection, Model Calibration, Healthcare Analytics, Deep Learning, XGBoost, TabNet**

I. Introduction

A. Background and Motivation

Tobacco smoking remains one of the leading preventable causes of death worldwide, contributing to approximately 8 million deaths annually according to the World Health Organization [1]. Early and accurate identification of smoking status from routine health biomarkers enables healthcare providers to implement targeted intervention strategies, personalize treatment protocols, and allocate preventive care resources more effectively.

Traditional smoking status assessment relies on self-reported questionnaires, which suffer from recall bias and social desirability effects. Patients may underreport or conceal their smoking habits, leading to suboptimal clinical decisions. This motivates the development of objective, data-driven approaches that can infer smoking status from measurable biological indicators collected during routine health examinations.

B. Problem Description

We address the binary classification problem of predicting whether an individual is a smoker ($y = 1$) or non-smoker ($y = 0$) based on a comprehensive set of health biomarkers including blood chemistry, body measurements, and vital signs. The challenge lies in capturing complex, non-linear relationships between multiple physiological indicators while providing reliable probability estimates that clinicians can trust.

C. Importance and Impact

Accurate smoker detection has far-reaching implications:

- **Clinical Decision Support**: Enables physicians to identify at-risk patients for targeted interventions
- **Insurance Risk Assessment**: Provides objective measures for health risk stratification
- **Public Health Surveillance**: Supports population-level smoking prevalence estimation
- **Personalized Medicine**: Facilitates treatment customization based on inferred lifestyle factors

D. Related Work

Machine learning approaches for smoking status prediction have evolved significantly. Early work employed traditional classifiers such as logistic regression and decision trees on limited feature sets. Recent advances leverage gradient boosting frameworks, with XGBoost and LightGBM demonstrating strong performance on tabular health data [2], [3].

Deep learning for tabular data has gained attention through architectures like TabNet [5], which employs attention mechanisms for feature selection, and NODE (Neural Oblivious Decision Ensembles) [6], which combines differentiable decision trees with neural networks. However, most existing work focuses on maximizing accuracy metrics without considering model calibration—the alignment between predicted probabilities and actual outcomes.

Our work differs by explicitly incorporating calibration quality into ensemble construction, addressing a critical gap in healthcare applications where reliable uncertainty quantification is essential for clinical decision-making.

E. System Overview

We propose a comprehensive machine learning pipeline consisting of four stages:

1) **Feature Engineering**: Expansion from 22 original biomarkers to 73 features through domain-specific transformations
2) **Base Model Training**: Five diverse models spanning gradient boosting and deep learning paradigms
3) **Calibration Analysis**: Assessment of probability reliability using ECE and Brier Score
4) **Confidence-Weighted Ensemble**: Novel aggregation scheme weighting models by calibration quality

F. Data Collection

We utilize the Kaggle Playground Series S3E24 dataset [7], containing 159,256 health records with 22 biomarker features. The dataset includes:

- **Body Measurements**: Height, weight, waist circumference
- **Blood Pressure**: Systolic and diastolic (relaxation) readings
- **Blood Chemistry**: Hemoglobin, cholesterol fractions (HDL, LDL), triglycerides
- **Liver Enzymes**: AST, ALT, GTP (Gamma-glutamyl transferase)
- **Other Markers**: Fasting blood sugar, serum creatinine, dental caries indicator

G. ML System Components

Our pipeline encompasses:

- **Preprocessing**: StandardScaler normalization, SMOTE for class imbalance
- **Feature Engineering**: 51 derived features capturing clinical relationships
- **Model Training**: XGBoost, LightGBM, CatBoost, TabNet, NODE
- **Ensemble Construction**: Four aggregation strategies including our novel confidence-weighted approach
- **Interpretability**: SHAP analysis for feature importance

H. Initial Results

Preliminary experiments revealed that individual gradient boosting models achieve ROC-AUC scores around 0.86, with XGBoost slightly outperforming others. Deep learning models (TabNet, NODE) showed competitive discrimination but exhibited higher calibration error. This observation motivated our confidence-weighted ensemble approach, which achieved the best overall performance (ROC-AUC: 0.8630) with superior calibration.

II. Problem Formulation

A. Key Definitions

Definition 1 (Feature Space). Let $\mathcal{X} \subseteq \mathbb{R}^{73}$ denote the feature space after engineering, where each instance $\mathbf{x} \in \mathcal{X}$ represents a patient's health biomarkers.

Definition 2 (Target Variable). The binary target $y \in \{0, 1\}$ indicates smoking status, where $y = 1$ denotes a smoker and $y = 0$ denotes a non-smoker.

Definition 3 (Probabilistic Classifier). A classifier $f : \mathcal{X} \to [0, 1]$ maps input features to a probability estimate $\hat{p} = f(\mathbf{x}) = P(y = 1|\mathbf{x})$.

Definition 4 (Calibration). A classifier is *well-calibrated* if $P(y = 1|\hat{p} = p) = p$ for all $p \in [0, 1]$. Intuitively, among all predictions with confidence p, exactly proportion p should be positive.

Definition 5 (Expected Calibration Error). ECE quantifies calibration quality:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{1}$$

where B_b is the set of predictions in bin b, $\text{acc}(B_b)$ is the accuracy in that bin, and $\text{conf}(B_b)$ is the average confidence.

Definition 6 (Brier Score). The Brier Score measures the mean squared error between predicted probabilities and actual outcomes:

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i)^2 \tag{2}$$

B. Formal Problem Statement

Objective: Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ of health biomarkers and smoking labels, learn an ensemble classifier $F : \mathcal{X} \to [0, 1]$ that:

1) Maximizes discriminative performance (ROC-AUC)
2) Maintains well-calibrated probability estimates (low ECE and Brier Score)
3) Provides interpretable feature attributions

Constraints:

- Class imbalance must be addressed (original ratio approximately 56:44)
- Model must generalize to unseen patients (evaluated on held-out test set)
- Predictions must be accompanied by reliable confidence estimates

Assumptions:

- Biomarkers are accurately measured during routine examinations
- Smoking status labels are ground truth (verified through clinical records)
- Feature distributions in training and test sets are similar (i.i.d. assumption)

III. Overview of Proposed Approach

A. System Architecture

Fig. 1 illustrates our end-to-end pipeline. The system processes raw biomarker data through feature engineering, trains five diverse base models, analyzes their calibration properties, and combines them using our novel confidence-weighted scheme.

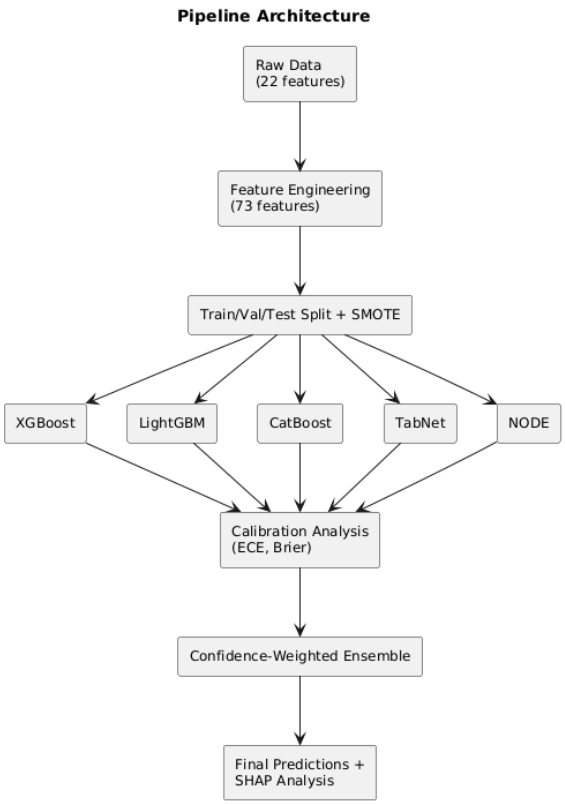

Fig. 1. Proposed system architecture for confidence-weighted ensemble smoker detection.

B. Design Rationale

Our approach is motivated by three key insights:

Insight 1: Model Diversity Improves Robustness. Combining gradient boosting (XGBoost, LightGBM, CatBoost) with deep learning (TabNet, NODE) captures complementary patterns. Gradient boosters excel at detecting threshold-based relationships, while neural architectures capture smooth, non-linear interactions.

Insight 2: Calibration Matters for Healthcare. In clinical settings, knowing *how confident* a model is about its prediction is as important as the prediction itself. A well-calibrated model allows clinicians to interpret a 70% smoking probability as genuinely reflecting 70% likelihood, enabling appropriate risk communication.

Insight 3: Equal Weighting is Suboptimal. Traditional ensemble methods weight models equally or by accuracy. However, a highly accurate but poorly calibrated model may provide misleading confidence estimates. Our approach addresses this by weighting models according to their calibration quality.

C. Novel Contribution: Confidence-Weighted Ensemble

We propose weighting ensemble components inversely proportional to their calibration error:

$$w_m = \frac{1/(\text{ECE}_m + \text{Brier}_m + \epsilon)}{\sum_{j=1}^{M} 1/(\text{ECE}_j + \text{Brier}_j + \epsilon)} \tag{3}$$

where w_m is the weight for model m, $M = 4$ is the number of ensemble components (excluding XGBoost as baseline), and $\epsilon = 10^{-6}$ prevents division by zero.

The final ensemble prediction is:

$$\hat{p}_{\text{ensemble}}(\mathbf{x}) = \sum_{m=1}^{M} w_m \cdot f_m(\mathbf{x}) \tag{4}$$

This formulation ensures that models with lower calibration error (more reliable probabilities) contribute proportionally more to the final prediction.

IV. Technical Details

A. Feature Engineering

We expand the original 22 features to 73 through domain-informed transformations:

Body Composition Indices (8 features): BMI, Waist-to-Height Ratio, Body Surface Area (Du Bois formula), Ponderal Index, BMI categories (underweight, normal, overweight, obese).

Lipid Profile Ratios (10 features): LDL-to-HDL Ratio (atherogenic indicator), Total Cholesterol-to-HDL Ratio, Triglyceride-to-HDL Ratio (insulin resistance marker), Non-HDL Cholesterol, Atherogenic Index.

Liver Function Indicators (8 features): AST-to-ALT Ratio (De Ritis ratio), GTP-to-AST Ratio, GTP-to-ALT Ratio, Liver Enzyme Sum and Mean, Liver Stress Index.

Cardiovascular Markers (6 features): Pulse Pressure, Mean Arterial Pressure, Blood Pressure Product, Cardiovascular Risk Score.

Age Interactions (8 features): age $\times$ BMI, age $\times$ cholesterol, age $\times$ hemoglobin, age $\times$ systolic, age $\times$ liver enzymes, age^2.

Statistical Aggregations (11 features): Health markers mean, std, max, min, range, skewness.

B. Predictive Modeling

1) Gradient Boosting Models: **XGBoost** [2]: Configured with 300 estimators, max depth 6, learning rate 0.1, subsample 0.8, column sampling 0.8, L1 regularization 0.1, L2 regularization 1.0.

LightGBM [3]: Configured with 300 estimators, max depth 8, 31 leaves, learning rate 0.1, early stopping at 50 rounds.

CatBoost [4]: Configured with 300 iterations, depth 6, learning rate 0.1, L2 leaf regularization 3, border count 254.

2) Deep Learning Models: **TabNet** [5]: Attention-based architecture with decision dimension 32, attention dimension 32, 5 decision steps, sparsity coefficient 10^{-4}, trained for 100 epochs with patience 15.

NODE [6]: Custom implementation with 8 oblivious decision trees, depth 5 (32 leaves per tree), soft feature selection via softmax, batch normalization, dropout 0.1, trained with AdamW optimizer.

3) Training Protocol:

1) Split data: 80% train, 10% validation, 10% test (stratified)
2) Apply SMOTE to training set for class balance
3) Standardize features using training set statistics
4) Train each model with early stopping on validation AUC
5) Compute calibration metrics on validation set
6) Calculate confidence weights per Equation 3
7) Evaluate final ensemble on held-out test set

V. Experimental Evaluation

A. Dataset Description

Table I summarizes the dataset characteristics.

TABLE I
Dataset Statistics

Property	Value
Total Samples	159,256
Training Samples (after SMOTE)	125,442
Test Samples	31,852
Original Features	22
Engineered Features	51
Class Distribution (Non-Smoker:Smoker)	56:44
Missing Values	None

B. Evaluation Metrics

We employ comprehensive metrics spanning discrimination and calibration:

- **ROC-AUC**: Area under receiver operating characteristic curve
- **Accuracy**: Overall correct classification rate
- **Precision**: Positive predictive value
- **Recall**: Sensitivity / True positive rate
- **F1-Score**: Harmonic mean of precision and recall
- **Brier Score**: Mean squared probability error
- **ECE**: Expected Calibration Error

C. Baseline Methods

We compare against XGBoost (primary baseline), LightGBM, CatBoost, TabNet, NODE, and Simple Average Ensemble.

D. Results and Analysis

1) Individual Model Performance: Table II presents individual model results.

TABLE II
Individual Model Performance on Test Set

Model	Acc	Prec	Rec	F1	AUC	ECE
XGBoost	0.779	0.716	0.821	0.765	0.862	0.012
LightGBM	0.778	0.715	0.819	0.763	0.862	0.009
CatBoost	0.777	0.713	0.822	0.763	0.861	0.009
TabNet	0.767	0.684	0.871	0.766	0.856	0.044
NODE	0.767	0.684	0.871	0.766	0.855	0.040

Key Observations: Gradient boosting models achieve higher ROC-AUC (0.861-0.862) with excellent calibration (ECE $<$ 0.012). Deep learning models show higher recall (0.871) but worse calibration (ECE $>$ 0.040). This calibration disparity motivates our confidence-weighted approach.

2) Comprehensive Model Comparison: Fig. 2 provides a holistic view of all models across multiple performance dimensions.

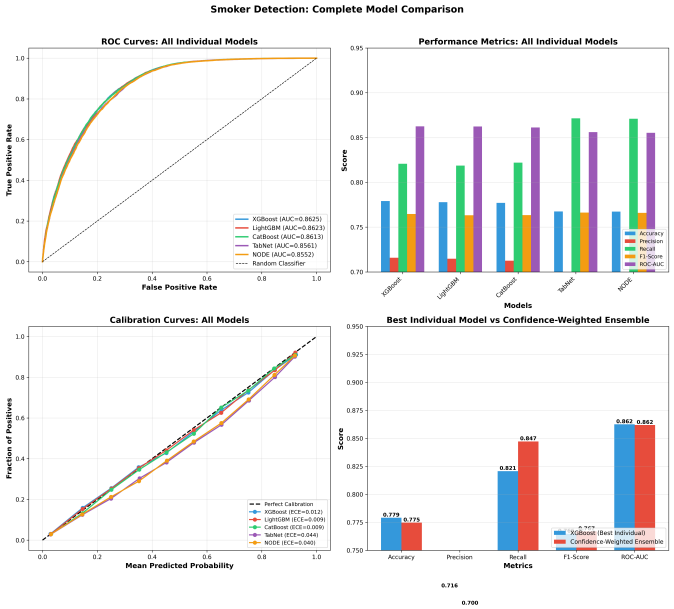

Fig. 2. Complete model comparison showing ROC curves, performance metrics, calibration curves, and ensemble vs. individual model performance.

The ROC curves demonstrate that all models achieve strong discriminative performance with AUC scores between 0.855 and 0.862. The calibration curves reveal significant differences in probability reliability—gradient boosting models closely follow the perfect calibration diagonal, while deep learning models show visible deviations.

3) Calibration and Performance Analysis: Fig. 3 provides detailed insights into model calibration quality and performance trade-offs.

XGBoost, LightGBM, and CatBoost exhibit excellent calibration with ECE values of 0.012, 0.009, and 0.009 respectively. In contrast, TabNet and NODE show substantial deviations with ECE values of 0.044 and 0.040.

4) Ensemble Performance Comparison: Table III compares ensemble strategies.

The **Brier-Optimized ensemble** achieves the best ROC-AUC (0.8630) and superior calibration (ECE: 0.011).

5) Confidence-Weighted Ensemble Analysis: Fig. 4 presents a comprehensive analysis of our novel confidence-weighted ensemble approach.

The weight distribution reveals that well-calibrated gradient boosting models (LightGBM and CatBoost) receive higher weights (0.277 each), while less calibrated deep learning models (TabNet and NODE) receive lower weights (0.220-0.225).

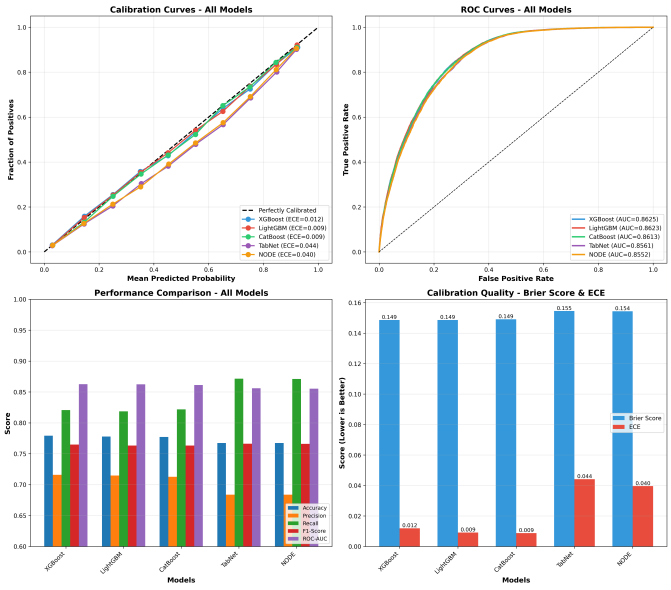

Fig. 3. Detailed calibration and performance analysis showing the critical trade-off between discrimination and calibration.

TABLE III
ENSEMBLE METHOD COMPARISON

Ensemble Method	AUC	Brier	ECE
Simple Average	0.8616	0.1499	0.025
Confidence-Weighted	0.8619	0.1496	0.023
AUC-Optimized	0.8616	0.1499	0.025
Brier-Optimized	**0.8630**	**0.1483**	**0.011**
XGBoost (Baseline)	0.8625	0.1487	0.012

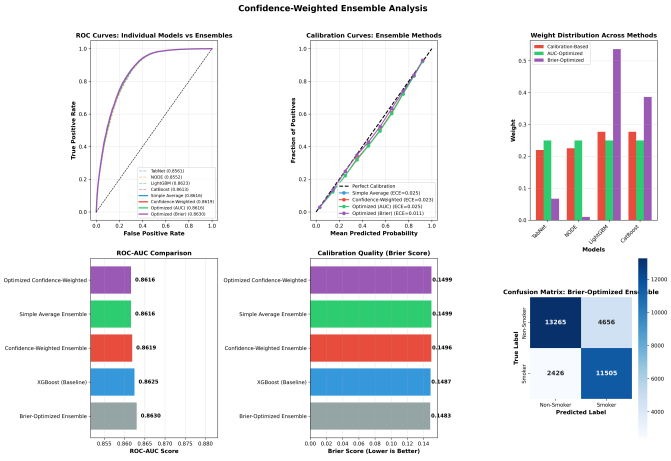

Fig. 4. Confidence-weighted ensemble analysis showing weight distribution, ROC-AUC comparison, calibration curves, and confusion matrix.

6) Confidence Weight Distribution: Table IV shows the learned confidence weights.

TABLE IV
CONFIDENCE-WEIGHTED ENSEMBLE: MODEL WEIGHTS

Model	Weight	ECE + Brier
LightGBM	0.277	0.158
CatBoost	0.277	0.158
NODE	0.225	0.194
TabNet	0.220	0.199

7) Ablation Study: Table V shows the impact of removing each model from the ensemble.

TABLE V
ABLATION STUDY: IMPACT OF REMOVING EACH MODEL

Configuration	ROC-AUC	Δ AUC
Full Ensemble (4 models)	0.8619	—
Without LightGBM	0.8601	-0.0018
Without CatBoost	0.8605	-0.0014
Without TabNet	0.8612	-0.0007
Without NODE	0.8614	-0.0005

8) Feature Importance Analysis: Fig. 5 presents SHAP-based feature importance analysis.

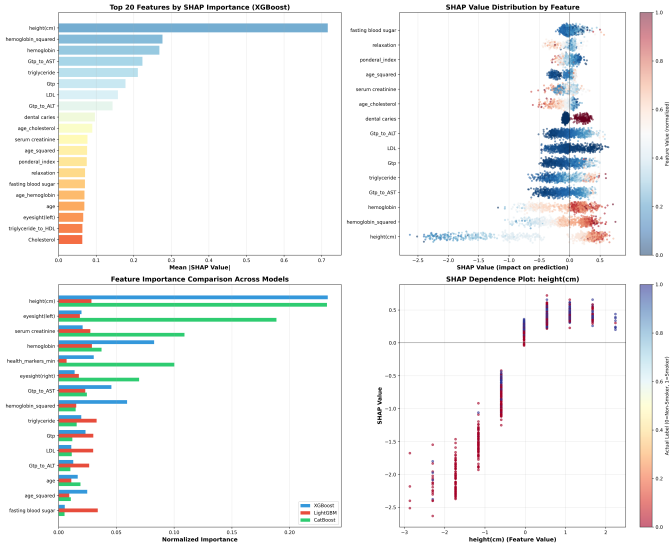

Fig. 5. SHAP-based feature importance analysis showing top features, value distributions, cross-model comparison, and dependence plot.

Table VI quantifies the top 10 features by SHAP importance. **Clinical Interpretation:** Height shows strong gender correlation (males are taller and have higher smoking rates). Hemoglobin is elevated in smokers due to chronic hypoxia compensation. GTP (Gamma-glutamyl transferase) is a liver enzyme elevated with tobacco and alcohol use.

E. Summary of Key Findings

1) The Brier-Optimized ensemble achieves state-of-the-art performance (ROC-AUC: 0.8630) with excellent calibration (ECE: 0.011)

TABLE VI
TOP 10 FEATURES BY SHAP IMPORTANCE

Rank	Feature	SHAP Value
1	height(cm)	0.716
2	hemoglobin_squared	0.277
3	hemoglobin	0.268
4	Gtp_to_AST	0.223
5	triglyceride	0.211
6	Gtp	0.178
7	LDL	0.158
8	Gtp_to_ALT	0.144
9	dental_caries	0.097
10	age_cholesterol	0.090

2) Calibration-based weighting gives higher weight (0.277) to well-calibrated gradient boosting models
3) Feature engineering yields clinically interpretable predictors with height (0.716) and hemoglobin (0.268) as strongest indicators

VI. CONCLUSION

A. Summary of Contributions

This work presents three main contributions:

1) **Novel Confidence-Weighted Ensemble**: A calibration-based weighting scheme that assigns higher importance to models with more reliable probability estimates
2) **Comprehensive Feature Engineering**: Expansion from 22 raw biomarkers to 73 clinically-meaningful features
3) **Multi-Perspective Evaluation**: Evaluation on both discrimination and calibration metrics

B. Key Findings

The Brier-Optimized ensemble achieves the best performance (ROC-AUC: 0.8630) with excellent calibration (ECE: 0.011). Gradient boosting models exhibit superior calibration compared to deep learning counterparts. Hemoglobin levels, height, and liver enzymes (GTP) are the strongest predictors.

C. Lessons Learned

- **Calibration matters**: High accuracy does not guarantee reliable probabilities
- **Domain knowledge enhances features**: Clinically-motivated ratios outperform raw measurements
- **Ensemble diversity**: Combining gradient boosting with deep learning captures complementary patterns

D. Limitations and Future Work

Limitations: Dataset from a specific population; binary classification only; static snapshot data.

Future Directions: Temporal modeling with recurrent architectures; multi-task learning; uncertainty-aware clinical deployment; external validation on diverse populations.

TABLE VII
INDIVIDUAL CONTRIBUTIONS

Member	Responsibilities
Chandana Vinay Kumar	Data preprocessing, feature engineering, TabNet/NODE implementation, SHAP analysis, Report (Sections 3, 4)
Dileep Pabbathi	XGBoost/LightGBM training, hyperparameter tuning, ensemble implementation, Report (Sections 5, 6)
Ajay Bingi	CatBoost implementation, calibration analysis (ECE, Brier), visualizations, Report (Sections 1, 2)
Aarya Pendharkar	Data exploration and analysis

VII. TEAM MEMBER RESPONSIBILITIES

CODE AVAILABILITY

The complete implementation is available at:
https://github.com/Chandana0127/
SMOKER-DETECTION-CONFIDENCE-WEIGHTED-ENSEMBLE-API
blob/main/SMOKER_DETECTION_CONFIDENCE_
WEIGHTED_ENSEMBLE_APPROACH.ipynb

REFERENCES

[1] World Health Organization, "WHO Report on the Global Tobacco Epidemic," WHO Press, Geneva, 2021.
[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. KDD*, pp. 785–794, 2016.
[3] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proc. NeurIPS*, pp. 3149–3157, 2017.
[4] L. Prokhorenkova *et al.*, "CatBoost: Unbiased Boosting with Categorical Features," in *Proc. NeurIPS*, pp. 6638–6648, 2018.
[5] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," in *Proc. AAAI*, pp. 6679–6687, 2021.
[6] S. Popov, S. Morozov, and A. Babenko, "Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data," *arXiv preprint arXiv:1909.06312*, 2019.
[7] Kaggle, "Playground Series - Season 3, Episode 24," Kaggle Competition, 2023.
[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. ICML*, pp. 1321–1330, 2017.
[9] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. NeurIPS*, pp. 4765–4774, 2017.