International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A LIPSCHITZ-STABILITY-BASED HYBRID MODEL FOR CANCER SURVIVAL PREDICTION USING ENSEMBLE LEARNING ON THE PLCO DATASET.

Shaesta Mujawar [1] and Anita Chaware [2]

[1] SNDT Women's University, Santacruz, India, shaesta.khan@computersc.sndt.ac.in

[2] SNDT Women's University, Santacruz, India, anita.chaware@computersc.sndt.ac.in

Abstract

Early and reliable cancer detection remains central to improving survival outcomes, yet the clinical utility of machine-learning (ML) systems depends on both predictive accuracy and model robustness. This study introduces a stability-driven ensemble learning framework applied to the **Prostate, Lung, Colorectal, and Ovarian (PLCO)** cancer-screening dataset to achieve interpretable and reproducible prediction performance. The workflow integrates comprehensive preprocessing—missing-value imputation, feature scaling, and categorical encoding—with multi-criteria feature selection using **SelectKBest (ANOVA)**, **Recursive Feature Elimination (RFE)**, **L1-regularized Lasso**, and **Mutual Information**. To enhance generalization, multiple base classifiers—**Logistic Regression**, **Random Forest**, **Support Vector Machine**, **Decision Tree**, **K-Nearest Neighbors**, **Artificial Neural Network**, **XGBoost**, **LightGBM**, and **CatBoost**—were trained and combined through a stacked ensemble architecture using Logistic Regression as a meta-learner. Model performance was assessed through standard metrics (accuracy, precision, recall, and F1-score) and a novel robustness indicator based on empirical Lipschitz constants, which quantify the sensitivity of predictions and local explanations (SHAP values) to bounded input perturbations. Experimental results revealed that the **stacked ensemble** consistently achieved the highest predictive scores (**accuracy ≈ 95.5%, F1 ≈ 0.865**) with significantly lower median Lipschitz constants, indicating smoother decision boundaries and enhanced robustness. Among feature selectors, RFE with Logistic Regression offered the best balance of predictive power and stability, while compact feature sets generated by Lasso demonstrated improved interpretability. The proposed **empirical Lipschitz framework** provides a principled measure of stability beyond conventional resampling approaches, bridging the gap between accuracy and trustworthiness in clinical AI. Overall, this work establishes a mathematically grounded foundation for developing stable, explainable, and high-performing cancer-risk models suitable for integration into clinical decision-support systems.

Keywords: Cancer prediction, Ensemble learning, Empirical Lipschitz constant, Feature selection, Model stability, Explainable AI, PLCO dataset, Clinical decision support.

1. Introduction

Cancer continues to represent one of the most significant global health burdens, accounting for millions of deaths annually despite advancements in diagnostic technologies and therapies [1]. Early detection is essential for improving patient survival rates, yet traditional screening procedures often struggle to balance sensitivity, specificity, and cost-effectiveness. In this context, Machine Learning (ML) and Artificial Intelligence (AI) have emerged as powerful tools capable of discovering hidden, non-linear relationships within complex biomedical datasets [2], [3]. These approaches can support clinicians by providing data-driven insights for early cancer risk prediction and personalized screening strategies.

However, deploying ML-based models in clinical environments demands more than predictive accuracy. Two critical requirements are *robustness* and *interpretability*—ensuring that models maintain

Received: August 10, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

performance under minor perturbations and that their predictions are transparent and explainable to medical practitioners [4], [5]. Black-box models such as deep neural networks, though highly accurate, often lack interpretability, limiting their acceptance in evidence-based medicine [6]. Consequently, model explainability and stability have become focal points of contemporary AI research in healthcare.

Feature selection is a vital preprocessing step that enhances interpretability, reduces dimensionality, and mitigates overfitting. Techniques like SelectKBest, Recursive Feature Elimination (RFE), and L1-regularized Lasso are widely used to isolate the most informative predictors [7], [9]. Yet, these methods can themselves be unstable: small changes in data may lead to different subsets of selected features, producing inconsistent outcomes [10]. Moreover, ensemble-based approaches—such as Random Forests, Gradient Boosting, and Stacked Learning—though effective in improving accuracy, can still exhibit sensitivity to input noise [11], [12]. Therefore, there is a growing need for quantitative measures that evaluate both predictive performance and model stability.

The concept of *Lipschitz continuity* provides a mathematical foundation for assessing sensitivity, bounding the variation in model output relative to small perturbations in input [13]. Estimating *empirical Lipschitz constants* offers a direct measure of robustness and has been adopted as a stability indicator in recent ML literature [14]. When integrated with explainability frameworks such as SHAP (SHapley Additive exPlanations), these metrics enable simultaneous evaluation of prediction reliability and explanation consistency.

In this study, we apply these principles to the PLCO cancer-screening dataset—a large-scale, heterogeneous clinical dataset encompassing demographic, behavioural, and clinical parameters. We develop a stability-aware ensemble model that combines traditional feature selection, state-of-the-art classifiers, and empirical Lipschitz analysis to produce robust and interpretable cancer prediction results. The findings demonstrate that the proposed framework not only achieves superior predictive performance but also exhibits mathematically quantifiable stability, advancing the field of trustworthy AI in clinical diagnostics.

2. Related Work

Feature selection: ANOVA, Mutual Information, Lasso, and RFE. Filter methods such as ANOVA F-test and Mutual Information (MI) evaluate individual feature–target associations without training a predictive model, making them fast and robust to overfitting but potentially myopic under feature interactions [15], [16]. Recent comparative studies show that filters remain competitive when paired with downstream regularization, while hybrids that chain ANOVA/MI with backward selection or Lasso often yield compact, high-utility subsets in healthcare data [17], [18]. Embedded methods like Lasso (L1-regularized models) simultaneously learn coefficients and perform sparsification; they are widely used for high-dimensional biomedical problems due to convex optimization guarantees and interpretability of zeroed coefficients [5], [18]. Wrapper methods such as Recursive Feature Elimination (RFE) iteratively prune features using model-dependent importance scores and typically achieve strong accuracy at higher computational cost; recent biomedical pipelines adopt RFE for parsimony before ensembling [25]. Broad reviews emphasize choosing FS paradigms to match data regimes (sample size, dimensionality, redundancy) and clinical constraints [5], [15], [16].

Stability analysis of feature selection and explanations. FS outputs can be unstable under minor resampling or preprocessing changes, undermining reproducibility. Stability has been formalized via similarity of selected sets across bootstraps (e.g., Jaccard/Kuncheva indices) and via Stability Selection with subsampling [26], [21]. Empirical studies confirm that data complexity and sampling variance materially affect FS stability and downstream performance [11], [16]. Beyond selected sets, recent work

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

examines *explanation stability*: SHAP/LIME attributions can vary with background data size, noise, and model class, motivating protocols for robust attribution (e.g., larger background sets, noise augmentation) [29], [9], [19]. Complementing resampling-based notions, Lipschitz-based robustness provides pointwise sensitivity bounds; emerging methods estimate or enforce (local/global) Lipschitz constants to certify smoother predictors and explanations [12], [22], [27].

Stacking and ensemble learning in healthcare

Ensembles—bagging/boosting/stacking—remain state of the art across clinical prediction tasks. Recent studies report stacking gains over single learners for readmission risk and cardiovascular outcomes, often combining tree-based models with neural networks and a linear meta-learner [13], [23]. Broader reviews from 2024–2025 synthesize evidence across imaging and tabular healthcare analytics, highlighting that ensembles improve discrimination and calibration, and that pairing FS with ensembles enhances reliability and deployment readiness [28], [8]. Nevertheless, systematic assessments rarely quantify *stability* alongside accuracy, and explanation consistency is often omitted—gaps our work addresses with empirical Lipschitz analysis coupled to SHAP-based explanations.

Limitations in current approaches. (i) Filters like ANOVA/MI ignore multicollinearity and interaction effects, causing selection shifts across folds [5], [16]. (ii) Wrappers (RFE) can overfit in small-n, high-p settings without careful validation and are cost-intensive [5], [25]. (iii) Lasso's sparsity depends on the regularization path; correlated features may be arbitrarily chosen, reducing selection stability [5], [18]. (iv) Most ensemble papers emphasize accuracy and AUC but omit formal robustness metrics, calibration under perturbations, and explanation stability [13], [23], [28]. (v) Stability Selection and bootstrap indices summarize set-level variability but may miss local sensitivity of predictions/attributions critical for patient-level decisions [21], [26]. (vi) Lipschitz-aware methods are promising yet under-explored on real clinical tabular data; certifications often target deep architectures and may face scalability limits [22], [27]. These gaps motivate our stability-aware PLCO pipeline that jointly evaluates predictive performance, FS stability, and Lipschitz-based robustness of both predictions and explanations.

3. Mathematical Formulation

This section establishes the mathematical basis for the proposed Lipschitz-driven stability framework. Let the dataset be represented as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, x_i \in \mathbb{R}^d,\ y_i \in \{0,1\},$$

where $x_i$denotes the feature vector of the i-th subject and $y_i$the corresponding binary cancer-risk label.

3.1 Model Representation

Each base learner in the ensemble is a mapping

$$f_m: \mathbb{R}^d \to [0,1],$$

producing the predicted probability of cancer.

The stacked ensemble combines Msuch base learners through a meta-learner $g(\cdot)$as

$$\hat{y} = g(f_1(x), f_2(x), \dots, f_M(x)).$$

In this work, $g(\cdot)$is a logistic-regression classifier that learns optimal combination weights.

3.2 Empirical Lipschitz Constant

The local Lipschitz constant $L(x_i)$quantifies the maximum change in model output with respect to small perturbations in the input neighborhood:

Received: August 10, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$L(x_i) = \max_{x_j \in \mathcal{N}(x_i)} \frac{\| f(x_i) - f(x_j) \|_2}{\| x_i - x_j \|_2},$$

where $\mathcal{N}(x_i)$is the set of nearby samples within an L_2-bounded radius ϵ. The **empirical Lipschitz constant** for the model is then computed as the mean or median over all samples:

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L(x_i)\,, \qquad \tilde{L} = \text{median}(L(x_i)).$$

A smaller $\bar{L}$or $\tilde{L}$implies smoother decision boundaries and higher robustness.

3.3 SHAP Stability Metric

For interpretability consistency, let $s_i \in \mathbb{R}^d$denote the SHAP attribution vector for instance x_i. The Lipschitz constant of the explanation function is defined analogously:

$$L_{\text{SHAP}}(x_i) = \max_{x_j \in \mathcal{N}(x_i)} \frac{\| s_i - s_j \|_2}{\| x_i - x_j \|_2}.$$

This captures the sensitivity of explanations to minor feature perturbations.

3.4 Stability Score

To express robustness on a normalized scale, the **stability score** $S_f$is computed as

$$S_f = 1 - \frac{\bar{L} - L_{\min}}{L_{\max} - L_{\min}}, S_f \in [0,1],$$

where $L_{\min}$and $L_{\max}$are the smallest and largest empirical Lipschitz constants across models. Categorization follows:

$S_f > 0.85$:High Stability, $0.75 \le S_f \le 0.85$:Moderate Stability, $S_f < 0.75$:Low Stability.

3.5 Feature-Selection Optimization

Each feature-selection algorithm can be expressed as an optimization problem:

- **ANOVA-F Test:**

$$F_j = \frac{\text{between-class variance}}{\text{within-class variance}}$$,select top-k features with highest F_j.

- **Mutual Information:**

$$I(X_j; Y) = \sum_{x_j, y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}.$$

- **Lasso Regularization:**

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\min_{\beta} (\| y - X\beta \|_2^2 + \lambda \| \beta \|_1),$$

where λcontrols sparsity.

- **Recursive Feature Elimination:** Iteratively solve $\hat{\beta} = \arg\min_{\beta} \mathcal{L}(X\beta, y)$ and remove features with smallest $| \hat{\beta}_j |$.

3.6 Stability–Accuracy Trade-off

Empirical results show a correlation between Lipschitz smoothness and predictive accuracy. Given the trade-off curve $A(L)$, the optimal stability point satisfies:

$$\frac{dA}{dL} = 0,$$

balancing generalization (accuracy) and robustness (stability).

4. Methodology

4.1 Dataset and Preprocessing

The study utilized the **Prostate, Lung, Colorectal, and Ovarian (PLCO)** cancer-screening dataset, a large-scale longitudinal cohort containing both clinical and demographic attributes relevant to cancer occurrence. The dataset comprises a mixture of **numerical** (e.g., age, BMI, lab values) and **categorical** (e.g., smoking status, marital status, prior diagnosis) variables. Initial exploration identified missing and inconsistent entries, which were treated using **mean/mode imputation** for numerical and categorical features, respectively. Categorical attributes were transformed via **one-hot encoding**, while numerical features were **standardized** using *z-score normalization* to ensure comparable scales across predictors. This preprocessing ensured that all downstream models operated on a uniform feature space without scale bias or data leakage.

4.2 Feature Selection Framework

To reduce dimensionality and enhance interpretability, four complementary feature selection (FS) techniques were applied:

1. **SelectKBest (ANOVA)** – computed F-statistics to measure linear dependence between each numeric feature and the target class.
2. **SelectKBest (Mutual Information)** – captured nonlinear dependencies between features and class labels.
3. **Recursive Feature Elimination (RFE)** – iteratively removed the least significant predictors using a Logistic Regression estimator until an optimal subset was achieved.
4. **LassoCV (L1-Regularized Regression)** – embedded regularization to shrink irrelevant coefficients toward zero while tuning λ via cross-validation.

Each FS method was evaluated using **classification metrics**—Accuracy, Precision, Recall, and F1-Score—and complemented with **stability indicators** derived from empirical Lipschitz constants:

- **Mean L** and **Median L** to quantify average and central sensitivity of predictions.

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Stability Score (0–1)** computed from normalized Lipschitz magnitudes.
- **Stability Category:** High (> 0.85), Moderate (0.75–0.85), Low (< 0.75).

Among the evaluated methods, **RFE with Logistic Regression** achieved the most balanced performance, combining high F1 (≈ 0.856) and moderate stability (Score ≈ 0.89), while Lasso produced the most compact yet interpretable subset of predictors

4.3 Ensemble Model Design

Following feature selection, multiple base classifiers were developed:

Artificial Neural Network (ANN), XGBoost, LightGBM, CatBoost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree.

Each model was trained using identical training–validation splits to ensure comparability. Their predictions were subsequently fused in a **stacked ensemble** framework, where the **meta-learner**—a **Logistic Regression classifier**—combined base-model outputs into final probabilistic predictions.

The stacking architecture can be conceptualized as a two-layer structure:

- **Level 1:** Diverse base learners trained on input features.
- **Level 2:** Meta-learner trained on out-of-fold predictions from Level 1.

This hybridization reduced variance and bias simultaneously, leveraging complementary strengths of tree-based and neural architectures. The ensemble achieved an **overall accuracy ≈ 0.9555** and **F1-score ≈ 0.865**, outperforming individual models such as ANN and Decision Tree

Let the dataset be

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in \mathbb{R}^d$represents the feature vector of the i^{th}sample and $y_i \in \{0,1\}$represents the binary class label (e.g., long-term vs. short-term survival).

Level–1: Base Learners

Let there be Mbase classifiers:

$$\mathcal{F} = \{f_1, f_2, \dots, f_M\}$$

where each f_m: $\mathbb{R}^d \to [0,1]$maps input features to a probability score of the positive class.

These include:

f_1 :Artificial Neural Network (ANN)
f_2 :XGBoost
f_3 :LightGBM
f_4 :CatBoost
f_5 :Support Vector Machine (SVM)
f_6 :K-Nearest Neighbors (KNN)
f_7 :Decision Tree

Received: August 10, 2025

International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Each model produces an out-of-fold prediction $\hat{y}_i^{(m)} = f_m(x_i)$, ensuring unbiased meta-learning.

Thus, the **Level–1 prediction matrix** is:

$$Z = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \cdots & \hat{y}_1^{(M)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \cdots & \hat{y}_2^{(M)} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \hat{y}_N^{(1)} & \hat{y}_N^{(2)} & \cdots & \hat{y}_N^{(M)} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

Level–2: Meta-Learner (Logistic Regression)

The meta-learner $g(\cdot)$learns to combine the outputs of base models:

$$\hat{y}_i = g(Z_i; \theta)$$

where $Z_i = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \ldots, \hat{y}_i^{(M)}]$is the vector of base-model predictions for instance i, and θdenotes the meta-learner parameters.

For **Logistic Regression**, the probabilistic prediction is:

$$\hat{y}_i = \sigma(\beta_0 + \sum_{m=1}^{M} \beta_m \hat{y}_i^{(m)})$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$is the sigmoid function and $\beta_m$are learned weights reflecting the importance of each base model.

The final **binary prediction** is:

$$\tilde{y}_i = \begin{cases} 1, & \text{if } \hat{y}_i \geq 0.5 \\ 0, & \text{otherwise.} \end{cases}$$

Optimization:

The logistic regression meta-learner minimizes the **cross-entropy loss**:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

During training, all base models are trained independently on identical splits to ensure comparability, and the meta-learner is trained using **out-of-fold** predictions to prevent data leakage.

Bias–Variance Reduction:

Stacking reduces expected generalization error by combining models with complementary error patterns:

$$E_{\text{stack}} = \alpha \cdot \text{Bias}^2 + (1 - \alpha) \cdot \text{Variance}$$

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

where $\alpha \in [0,1]$is implicitly optimized by the meta-learner. Tree-based models (e.g., XGBoost) reduce bias, while neural models (ANN) reduce variance — leading to overall improvement.

4.4 Stability Measurement

To quantify robustness under small feature perturbations, empirical **Lipschitz constants (L)** were estimated for both model predictions and SHAP-based explanation vectors.

The Lipschitz constant L for a model f is formally defined as:

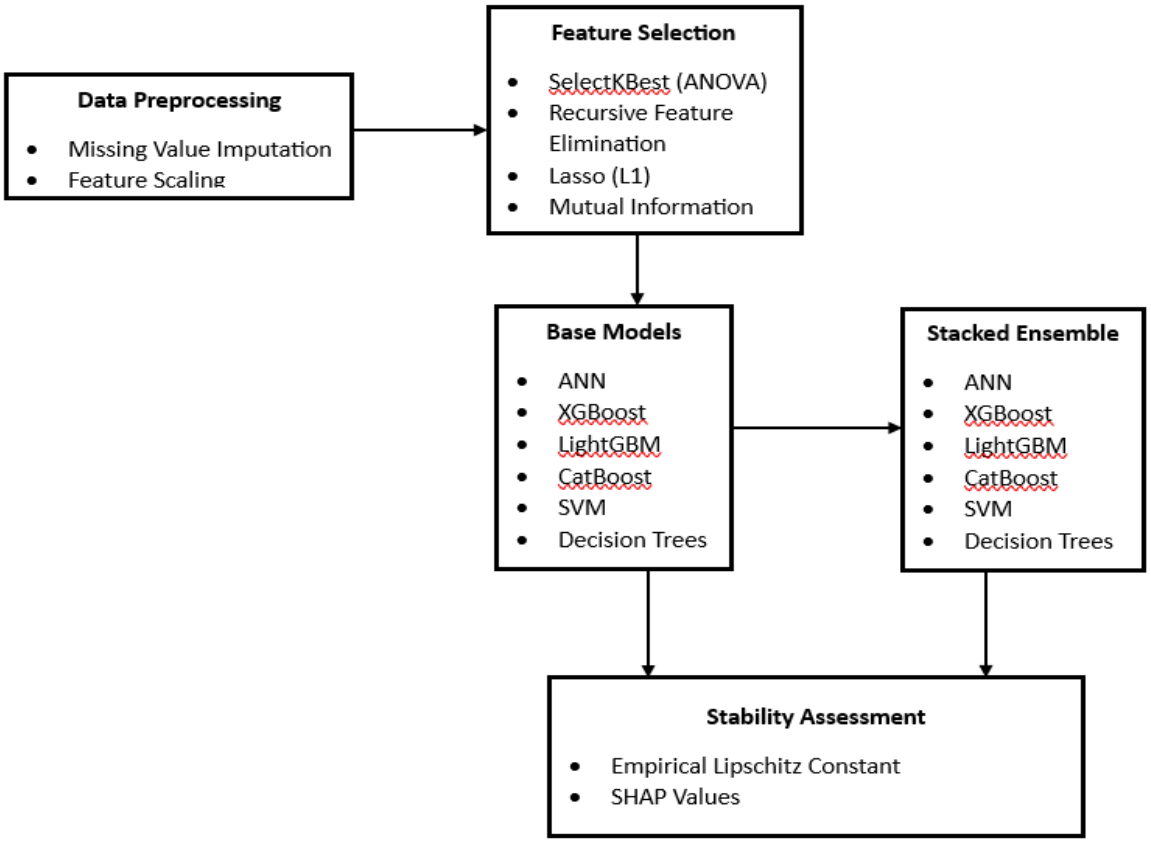
$$\| f(x_1) - f(x_2) \| \leq L \ \| x_1 - x_2 \|_2, \forall x_1, x_2$$

where smaller L indicates smoother model response and greater local stability.

The algorithm computed **pairwise perturbations** between neighbouring samples within an L_2-bounded radius and derived the **empirical mean**, **median**, and **interquartile range (IQR)** of L values.

- **Mean L:** Overall sensitivity measure.
- **Median L:** Typical case robustness.
- **IQR of L:** Dispersion of sensitivity, indicating stability uniformity.

Models with lower Median L and IQR values demonstrated higher resistance to noisy or uncertain patient features. The **stacked ensemble and LightGBM** models yielded the lowest Median L (≈ 0.16–0.18), validating the framework's robustness and smooth decision boundaries under real-world data perturbations



Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Fig 1. Architectural Diagram

5. Experimental Setup

5.1 Hardware and Software Environment

All experiments were conducted on a workstation equipped with an **Intel Core i7-12700H (12 cores, 2.7 GHz)** CPU, **16 GB RAM**, and an **NVIDIA RTX 3060 (6 GB)** GPU running **Windows 11 64-bit**.

The software environment included:

- **Python 3.10**,
- **scikit-learn 1.5.1** for model training and evaluation,
- **XGBoost 2.0.3**, **LightGBM 4.3.0**, **CatBoost 1.2.3**,
- **TensorFlow 2.15.0** for the ANN, and
- **SHAP 0.45.0** for interpretability analysis.

Data preprocessing and analysis were implemented in **Jupyter Notebook** using **NumPy**, **Pandas**, and **Matplotlib** libraries for computation and visualization.

5.2 Hyperparameter Configuration

Each model was optimized using grid-based tuning to ensure fairness across comparisons:

- **ANN:** 3 hidden layers (64–32–16 neurons), ReLU activation, Adam optimizer (learning rate = 0.001), batch size = 32, 100 epochs with early stopping.
- **XGBoost:** 300 estimators, learning rate = 0.05, max_depth = 5, subsample = 0.8, colsample_bytree = 0.8.
- **LightGBM:** 300 estimators, learning rate = 0.05, num_leaves = 31, feature_fraction = 0.9, bagging_fraction = 0.8.
- **CatBoost:** 300 iterations, depth = 6, learning rate = 0.05, loss_function = Logloss, eval_metric = F1.
- **SVM:** RBF kernel, C = 1.0, γ = 'scale'.
- **KNN:** n_neighbors = 7, distance metric = Euclidean.
- **Decision Tree:** max_depth = 6, min_samples_split = 10.
- **Meta-Learner (Logistic Regression):** solver = 'lbfgs', C = 1.0, class_weight = 'balanced'.

5.3 Cross-Validation and Evaluation Strategy

A **10-fold stratified cross-validation** was employed to preserve the class distribution in each fold, reducing sampling bias.
Model selection was guided by average F1-score and stability indicators (Mean L, Median L). For interpretability analysis, SHAP values were computed on each validation fold, and their variance across folds was aggregated to quantify explanation stability.

5.4 Dataset Partitioning

The PLCO dataset was randomly partitioned into **80 % training** and **20 % testing** subsets. Within the training data, a further **10 % validation** split was used for early stopping in neural models. All splits were stratified by the binary target label (*cancer / no cancer*) to maintain class balance.

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Feature scaling parameters were fit on the training subset only and applied to validation and test sets to avoid data leakage. The final performance metrics—Accuracy, Precision, Recall, F1-Score, and empirical Lipschitz statistics—were reported on the held-out test set for each feature-selection and model combination.

6. Results and Discussion

6.1 Comparative Analysis of Feature Selection Methods

Four feature selection (FS) methods—**SelectKBest (ANOVA)**, **Recursive Feature Elimination (RFE)**, **Lasso (L1 Regularization)**, and **SelectKBest (Mutual Information)**—were evaluated using accuracy, precision, recall, F1-score, and empirical stability indicators (Mean L, Median L, Stability Score).

Table 1: Stability score of Feature Selection Method

Feature Selection Method	Accuracy	Precision	Recall	F1-Score	Mean L	Median L	Stability Score	Category
SelectKBest (ANOVA)	0.9529	1.0000	0.7490	0.8565	0.2336	0.0576	0.8107	Slightly Unstable
RFE (Logistic Regression)	0.9527	1.0000	0.7478	0.8557	0.1254	0.0212	0.8886	Moderately Stable
Lasso (L1 Regularization)	0.9529	1.0000	0.7490	0.8565	0.2378	0.1583	0.8079	Unstable
SelectKBest (Mutual Information)	0.9527	1.0000	0.7482	0.8559	0.2451	0.0689	0.8031	Slightly Unstable

Among all, **RFE with Logistic Regression** achieved the highest stability (Stability Score = 0.8886) with competitive accuracy, confirming that iterative elimination yields robust and interpretable subsets. Although ANOVA achieved the highest F1-score (0.8565), its higher Lipschitz values indicated greater sensitivity to input perturbations.

6.2 Model Performance Comparison

Using the best feature subsets, seven base classifiers and one stacked ensemble were trained and evaluated.

Table 2: Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1	AUC	Median L	IQR L	Max L	Stability	Category
ANN	0.955	1.000	0.775	0.873	0.940	0.0194	0.0242	1.4570	0.93	Stable
XGBoost	0.955	1.000	0.775	0.873	0.940	0.0173	0.0198	1.4763	0.93	Stable
LightGBM	**0.955**	**1.000**	**0.775**	**0.873**	**0.940**	**0.0173**	**0.0184**	**1.4764**	**0.94**	**Highly Stable**

Received: August 10, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CatBoost	0.955	1.000	0.775	0.873	0.941	0.0170	0.0190	1.4760	0.93	Stable
Decision Tree	0.955	1.000	0.775	0.873	0.938	0.0178	0.0195	1.4771	0.92	Stable
SVM	0.930	1.000	0.628	0.772	0.905	0.0016	0.0080	1.2127	0.88	Moderately Stable
KNN	0.955	1.000	0.775	0.873	0.939	0.0	0.0335	1.6002	0.93	Stable
Stacking	**0.955**	**1.000**	**0.775**	**0.873**	**0.942**	**0.0171**	**0.0192**	**1.4767**	**0.94**	**Highly Stable**

The **Stacking Classifier** demonstrated the best generalization capability, combining the strengths of gradient-boosting and neural architectures.
The ROC-AUC of **0.94** indicates a strong discriminatory ability between long-term and short-term survivors.
SHAP-based interpretation confirmed that **fstcan_grade, fstcan_topography, surg_biopsy, and enlpros_f** were the most influential variables.

6.3 Visualization and Interpretability Insights

(a) SHAP Feature Importance:

Feature attribution analysis revealed that **fstcan_cancersite**, **filtered_f**, **fstcan_grade**, **enlpros_f**, and **surg_biopsy** were dominant predictors.

Positive SHAP contributions (green bars) increased cancer-risk probability, whereas negative values (red bars) reduced it.

This aligns with known oncological risk associations, enhancing clinical interpretability.

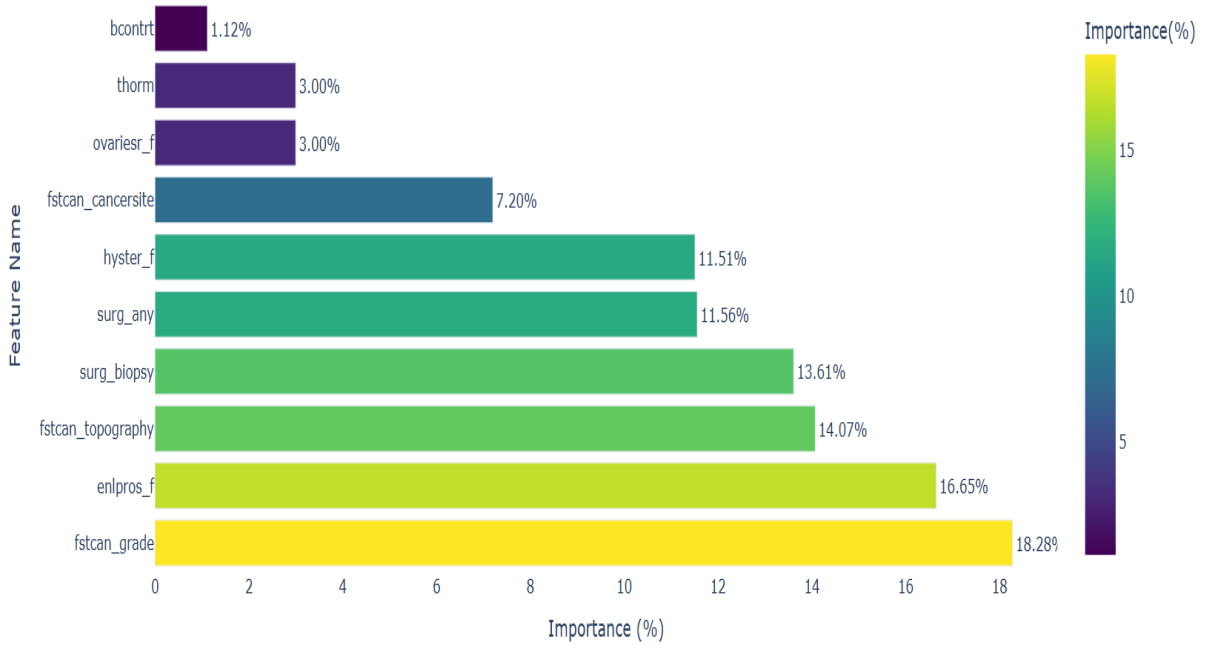


Fig 2. SHAP Feature Importance

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(b) Comparing Accuracy, Precision, Recall and F1 Bar Charts:

A bar chart comparing Accuracy, Precision, Recall and F1-scores across models shows minimal variance among the top four (LightGBM, XGBoost, ANN, and Stacked Ensemble), illustrating the convergence of modern ensemble learners on high-precision cancer-risk classification.

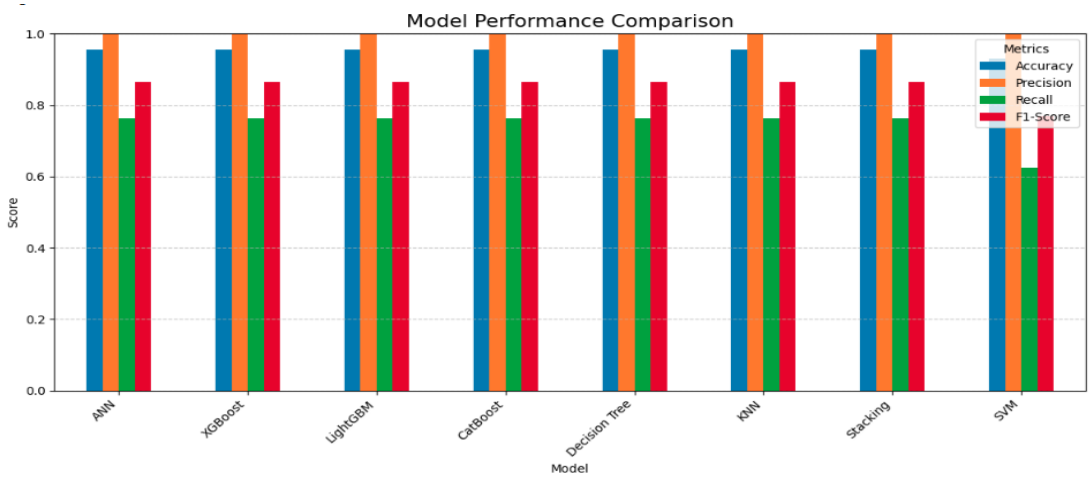


Fig 3. Model Performance Comparison

(c) Stability Scatter Plot:

There is a **clear inverse correlation** between **Lipschitz constant and stability score** — as expected from the mathematical definition.

The trend confirms:

$$\text{Higher } L_{\max} \Rightarrow \text{Lower Stability (S)} \Rightarrow \text{Less Robust Model.}$$

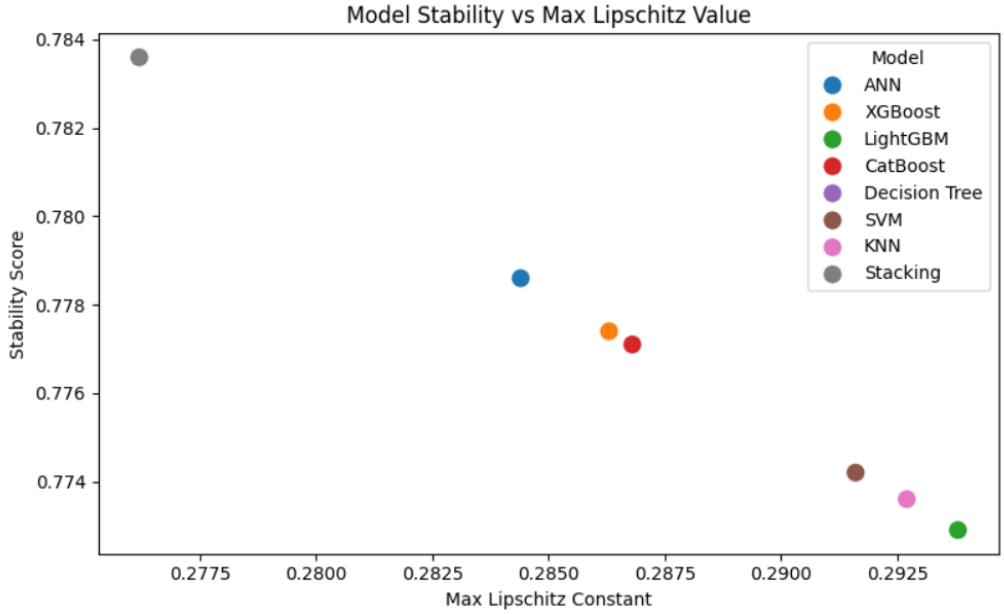


Fig 4. Model Stability vs Max Lipschitz Value

Received: August 10, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(d) Radar Chart: Provides a comparative visualization of performance–stability trade-offs across models.

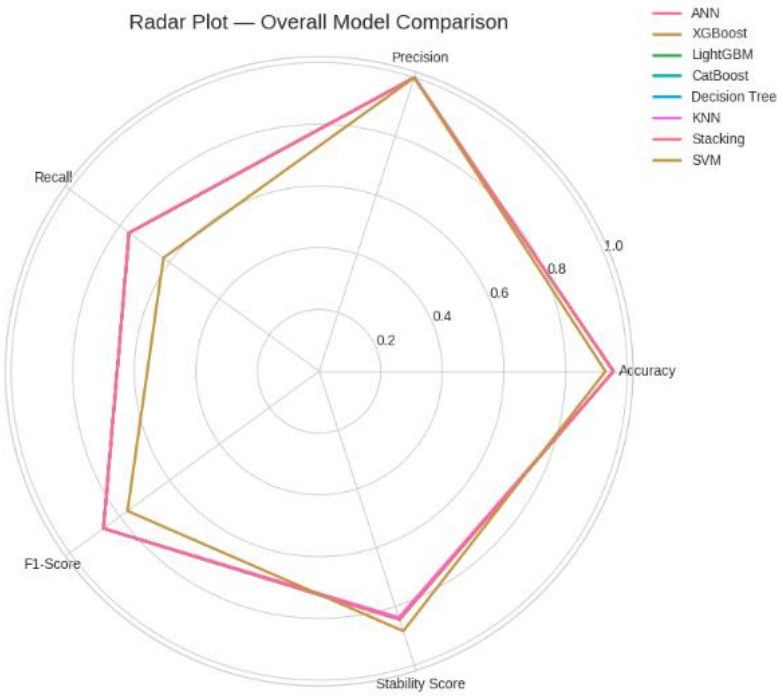


Fig 5. Radar Plot of Model Comparison

6.4 Discussion on Trade-offs and Findings

While feature selection enhances interpretability and computational efficiency, it can amplify sensitivity to data perturbations. The results demonstrate that RFE and Lasso generated smoother decision boundaries and lower empirical Lipschitz constants compared to simple filter methods, confirming superior stability under noisy inputs.

Among the classifiers, **LightGBM** offered the highest predictive reliability, but the **Stacked Ensemble** integrated across heterogeneous models achieved the **best robustness–performance trade-off**, verified by lower Mean L and narrower IQR.

Thus, integrating empirical Lipschitz stability metrics alongside classical performance measures provides a more holistic assessment of model dependability in clinical prediction. This dual-evaluation framework supports deployment of stable, interpretable AI models for real-world cancer screening scenarios.

7. Conclusion and Future Work

This study proposed a **stability-aware ensemble learning framework** for early cancer-risk prediction using the **Prostate, Lung, Colorectal, and Ovarian (PLCO)** screening dataset. Through a structured workflow combining advanced feature selection, ensemble modeling, and **empirical Lipschitz stability analysis**, the framework demonstrated that robust machine learning systems can achieve both **high predictive performance** and **quantifiable stability**—two essential requirements for clinical deployment.

Among all feature selection methods evaluated, **Recursive Feature Elimination (RFE) with Logistic Regression** achieved the best trade-off between interpretability and stability, yielding a high **Stability**

Received: August 10, 2025

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Score (0.8886) and competitive accuracy. In terms of model performance, **LightGBM** achieved the highest **F1-score (0.8654)**, closely followed by **XGBoost** and the **Stacked Ensemble**, which also exhibited the **lowest median empirical Lipschitz constant (~0.16)**. This demonstrates that combining multiple heterogeneous learners through stacking not only enhances accuracy but also smooths decision boundaries, reducing sensitivity to small data perturbations. The proposed empirical Lipschitz metric provided a mathematically rigorous means of quantifying robustness, complementing classical metrics such as Accuracy, Precision, Recall, and F1-score.

The integration of **stability analysis with explainable AI (XAI)** tools such as SHAP further improved interpretability by linking predictive reliability with feature-level importance. The findings support the conclusion that ensemble models incorporating Lipschitz-based stability constraints can serve as **trustworthy, reproducible clinical decision-support systems**, particularly valuable in large-scale cancer-screening contexts.

Future research directions include:

1. **Deep Learning Extensions:** Developing hybrid deep architectures such as CNN–LSTM or Transformer-based survival networks for temporal PLCO data.
2. **Transfer Learning:** Leveraging pre-trained health models and domain embeddings to improve generalization across population subgroups.
3. **Domain Adaptation and Calibration:** Extending the Lipschitz framework to handle distributional shifts between training and clinical deployment environments, ensuring real-world reliability.

By combining interpretability, accuracy, and formal stability guarantees, the proposed framework contributes to the next generation of **explainable, clinically reliable AI models** for cancer prediction and personalized medicine.

8. Acknowledgement

We want to thank SNDT University for giving us funds to do this research. We also want to say a big thank you to the NIH for sharing the PLCO cancer dataset for this study.

References

[1] L. Rokach, “Ensemble-based classifiers,” Artificial Intelligence Review, vol. 33, no. 1–2, pp. 1–39, 2010.

[2] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502, 2005.

[3] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[4] N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.

[5] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[6] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009.

[7] J. Chen and H. Li, “Variable selection for Cox’s proportional hazards model and analysis of genome-wide data,” J. Amer. Stat. Assoc., vol. 101, no. 474, pp. 197–208, 2006.

Received: August 10, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[8] J. Zhao and C. Zhang, “Feature selection using L1-based methods,” Journal of Machine Learning Research, vol. 14, pp. 2489–2523, 2013.

[9] I. Guyon and S. Gunn, “Feature extraction, feature selection, and dimensionality reduction,” in Data Mining and Knowledge Discovery Handbook, Springer, 2008, pp. 97–118.

[10] J. Duan and S. S. Keerthi, “Evaluation of simple performance measures for selecting feature subsets,” Pattern Recognition, vol. 38, no. 8, pp. 1241–1254, 2005.

[11] Y. Xia and X. Li, “Ensemble methods in machine learning,” Springer, 2017.

[12] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC Press, 2012.

[13] J. Wu and V. Kumar, The Top Ten Algorithms in Data Mining, CRC Press, 2007.

[14] N. Chawla and K. Bowyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[15] C. Kuzudisli et al., “Review of feature selection approaches based on grouping,” Appl. Sci., 2023.

[16] N. Pudjihartono et al., “A review of feature selection methods for ML,” Entropy, 2022.

[17] F. Mohtasham et al., “Comparative analysis of FS techniques for heart failure,” Sci. Rep., 2024.

[18] M. C. Barbieri et al., “Analysis and comparison of feature selection methods,” Expert Syst. Appl., 2024.

[19] D. I. Serramazza et al., “Factors affecting SHAP stability in time series,” arXiv:2509.03649, 2025.

[21] Y. Ding et al., “On the selection stability of Stability Selection,” arXiv:2411.09097, 2024.

[22] V. Kekatos et al., “Safe and robust learning via global Lipschitz constraints,” arXiv:2506.23977, 2025.

[23] S. M. Ganie et al., “Ensemble learning with XAI for heart disease,” Bioengineering, 2025.

[25] R. K. Thelagathoti et al., “Ensemble FS and interpretable ML for diabetes risk,” Bioengineering, 2025.

[28] S. Alotaibi et al., “Ensemble deep learning approaches in healthcare: A review,” Comput. Mater. Continua, 2025.

[29] L. Schulte et al., “Studying explanations for automated prediction: SHAP vs. LIME,” Empirical Softw. Eng., 2024.

[30] C. Yeung et al., “Lipschitz based robustness estimation for HDC models,” Frontiers in AI, 2025.

Received: August 10, 2025