Machine Learning and Stacked Logistic Regression for Cancer Risk Prediction on a Public Lifestyle Dataset

Qinglan Ouyang
qinglanm@umich.edu

Weiying Sun
weiyings@umich.edu

Jiadong Zhu
jiadongz@umich.edu

[GitHub Repository]

Abstract

Cancer prediction using demographic and lifestyle factors is important when detailed clinical data are unavailable. We compare seven classification models on a public cancer prediction dataset and investigate a stacked logistic regression approach combining random forest and gradient boosting. Previous work has mainly compared individual models, often showing strong performance from tree-based ensembles, but rarely addresses class imbalance or explores model combination. Our goal is to assess whether stacking improves predictive performance, particularly F1-score and recall, while maintaining interpretability.

1 Introduction

Cancer remains one of the leading causes of morbidity and mortality worldwide, and early identification of individuals at elevated risk is essential for prevention and timely intervention (Kourou et al., 2015). In many real-world settings, comprehensive clinical or genomic data may not be available, and risk assessment must rely on routinely collected demographic and lifestyle variables such as age, body mass index (BMI), smoking status, alcohol intake, physical activity, genetic risk, and cancer history. Understanding how well these predictors can distinguish individuals with and without cancer is of both scientific and practical relevance.

The public "Cancer Prediction" dataset used in this project has been widely analyzed in previous machine learning studies. Most prior work evaluates the performance of individual models—such as logistic regression, support vector machines, random forest, or gradient boosting—on this dataset, often reporting strong accuracy for tree-based ensemble methods (Caruana and Niculescu-Mizil, 2006). However, these analyses typically evaluate models in isolation and do not explore how different learners might be combined. Moreover, many studies emphasize overall accuracy despite mild class imbalance, even though metrics such as recall and F1-score carry more clinical importance due to the cost of missed cancer cases (Saito and Rehmsmeier, 2015).

This project has two primary goals. The first is to systematically compare classical statistical models and modern machine learning methods for binary cancer prediction under a consistent cross-validation framework. The second is to investigate whether combining strong ensemble models through a stacked logistic regression meta-model can improve predictive performance while retaining interpretability. Stacking allows non-linear models, such as random forest and gradient boosting, to capture complex relationships in the data, while logistic regression provides a transparent final layer that summarizes their contributions. We evaluate models using accuracy, precision, recall, and F1-score on a held-out test set, with parallel cross-validation to reduce computational time.

2 Data

The dataset comes from the publicly released "Cancer Prediction Dataset" on Kaggle, a synthetic dataset designed to mimic plausible risk relationships from epidemiological studies. It contains 1,500 individuals in a case–control structure, including cancer cases and non-cancer controls. This allows us to study risk differences and build supervised learning models for binary cancer prediction. Predictors include four continuous variables (Age, BMI, PhysicalActivity, AlcoholIntake), three binary variables (Gender, Smoking, CancerHistory), and one three-level categorical variable (GeneticRisk). The outcome variable, Diagnosis, indicates cancer status.

According to the initial exploratory analysis. The outcome variables were slightly imbalanced (approximately 60% non-cancer, 40% cancer) (Figure 1a), supporting the use of metrics beyond overall accuracy. Exploration also suggests that some predictors are in a non-linear trends. For example, continuous variables such as BMI and physical activity appears to vary across different value ranges, not increasing proportionally. These indicate that flexible ML models and interpretability tools are needed to better suit for capturing complex risk patterns.

3 Methods

All categorical variables were converted to factors: Gender (Male/Female), Smoking (No/Yes), CancerHistory (No/Yes), GeneticRisk (Low/Medium/High), and Diagnosis (NoCancer/Cancer). Continuous variables (Age, BMI, PhysicalActivity, AlcoholIntake) were standardized using z-score normalization, with parameters learned from the training set and applied to both training and test sets to prevent data leakage. The dataset was split into training (80%) and test (20%) sets using stratified random sampling based on the outcome variable to maintain class balance across splits. A fixed random seed (42) was used to ensure reproducibility.

We compared seven classification algorithms: logistic regression (glm), decision tree (rpart), random forest (rf), gradient boosting (gbm), support vector machine with radial basis function kernel (svmRadial), k-nearest neighbors (knn), and naive Bayes (nb). All models were trained using the `caret` package in R with 5-fold stratified cross-validation for hyperparameter tuning. For decision trees, we tuned the complexity parameter (cp) over a grid from 0.01 to 0.1. Other models used their default tuning grids. Model selection was based on accuracy, though we also tracked precision, recall, and F1-score given the mild class imbalance. To reduce computational time, parallel processing was enabled using the `doParallel` package, utilizing all available CPU cores minus one. All models were evaluated on the held-out test set using accuracy, precision, recall, and F1-score, with “Cancer” designated as the positive class.

We implemented a two-level stacking approach to combine predictions from the best-performing ensemble models. In the first level, we selected random forest and gradient boosting as base learners based on their strong individual performance. To generate out-of-fold (OOF) predictions without data leakage, we performed 5-fold cross-validation on the training set: for each fold, we trained random forest and gradient boosting models on the training portion using 3-fold cross-validation for hyperparameter tuning, and obtained probability predictions for the validation portion. This process yielded OOF probability predictions for all training samples from both base learners. These OOF predictions were then used as features for a second-level meta-learner. We trained a logistic regression model with LASSO regularization (L1 penalty) using `glmnet`, with the regularization parameter λ selected via 5-fold cross-validation to minimize deviance. The LASSO penalty helps prevent overfitting and provides interpretable coefficients indicating the relative contribution of each base learner. For final predictions on the test set, we retrained the random forest and gradient boosting models on the full training set, obtained their probability predictions on the test set, and fed these into the trained meta-learner to produce final stacked predictions.

4 Results

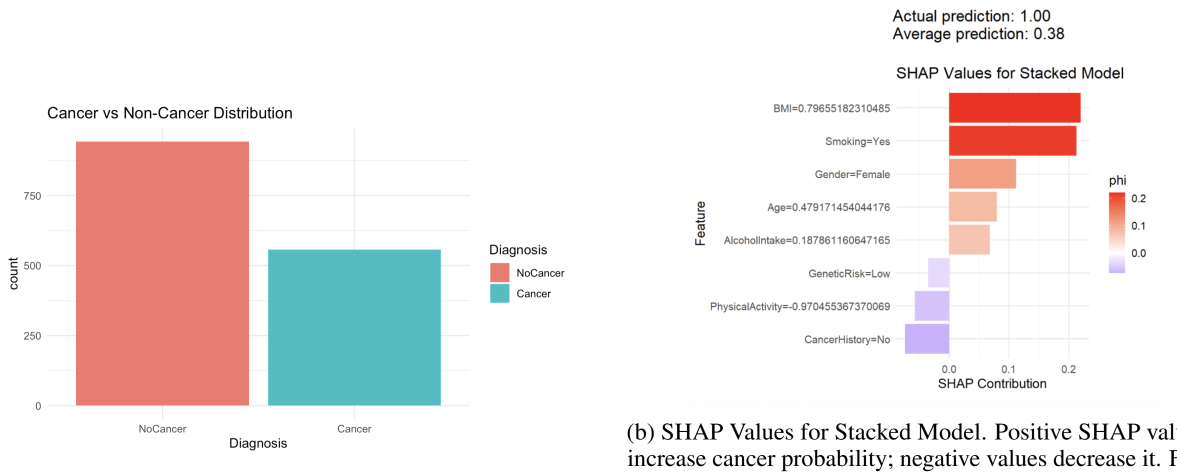
4.1 Individual Model Performance

Table 1 summarizes the test set performance for all seven models.

Tree-based ensemble methods achieved the highest performance, with gradient boosting achieving the best test set accuracy (0.953), precision (0.971), recall (0.901), and F1-score (0.935). Random forest ranked second (accuracy: 0.923, F1: 0.894). Logistic regression and SVM showed comparable performance (accuracy: 0.883), while other models performed lower, likely due to limited capacity to capture complex interactions or sensitivity to feature scaling and class imbalance.

4.2 Stacked Model Performance

The stacked logistic regression model achieved test set accuracy of 0.943, precision of 0.963, recall of 0.948, and F1-score of 0.955. Compared to the baseline logistic regression model, the stacked model achieved substantially higher accuracy (+6.8%), precision (+12.2%), recall (+15.6%), and F1-score (+13.8%), which is particularly important given the clinical cost of missed cancer cases. The meta-learner coefficients showed gradient boosting received a much higher weight (10.92) than random forest (0.64). The test AUC was 0.949 (training: 0.962), compared to 0.926 for baseline logistic regression, suggesting minimal overfitting



(a) Cancer vs Non-Cancer Distribution

(b) SHAP Values for Stacked Model. Positive SHAP values increase cancer probability; negative values decrease it. Features are ordered by importance.

Figure 1: Exploratory visualizations comparing outcome balance and stacked model.

Model	Test Accuracy	Precision	Recall	F1-Score
Gradient Boosting (gbm)	0.953	0.971	0.901	0.935
Random Forest (rf)	0.923	0.915	0.874	0.894
Logistic Regression (glm)	0.883	0.858	0.820	0.839
SVM (svmRadial)	0.883	0.880	0.793	0.834
Decision Tree (rpart)	0.876	0.856	0.802	0.828
k-NN (knn)	0.819	0.913	0.568	0.700
Naive Bayes (nb)	0.789	0.843	0.532	0.652
Stacked Model	0.943	0.963	0.948	0.955

Table 1: Test set performance metrics for all seven classification models and the stacked model.

and good generalization. Figure 2 compares the ROC curves.

4.3 Model Diagnostics

Calibration analysis showed the stacked model's predicted probabilities aligned well with observed event rates (Brier score: 0.053), much smaller than the baseline prediction of 0.25. Figure 3 displays the calibration plot.

Residual diagnostics indicated residuals were approximately randomly distributed around zero, with no strong patterns suggesting systematic model misspecification. Figure 4 shows the deviance residuals versus fitted probabilities and Q-Q plot of standardized residuals.

Feature importance analysis revealed cancer history, BMI, alcohol intake, and high genetic risk as the most important predictors, consistent with established epidemiological risk factors. Table 2 displays feature importance rankings from both base learners.

Partial dependence plots for BMI and alcohol intake showed non-linear relationships with cancer risk, justifying the use of ensemble methods. Figure 5 displays these plots.

5 Discussion

The stacked model demonstrated strong performance with improved recall and F1-score compared to individual models, which is crucial for cancer risk assessment where missing high-risk individuals has severe clinical consequences. The model achieved good probability calibration (Brier score: 0.053) and showed minimal overfitting (train-test AUC gap: 0.013), indicating good generalization ability.

However, stacking performance remains dependent on base learner diversity. The two ensemble methods used may have learned similar patterns, thus limiting the additional benefits of stacking. To further improve model performance, we fine-tuned the RF and GBM before stacking. However, experimental results show that this approach did not significantly improve the performance of the stacked model; in fact, it exacerbated overfitting on the training set in some cases. This indicates that overfitting cannot be solved simply by optimizing the base models or adjusting the regularization parameters. The fine-tuning process may reinforce specific patterns already learned in the training data, which may be further amplified during stacking.

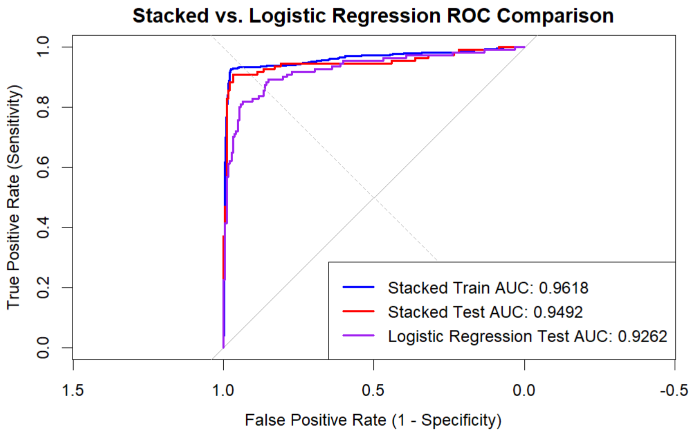


Figure 2: ROC curve comparison for the stacked model (training and test sets) and baseline logistic regression model.

Feature	RF Importance	GBM Importance
Cancer History (Yes)	100.00	20.91
BMI	83.93	10.69
AlcoholIntake	80.48	11.10
Genetic Risk (High)	79.50	17.86
PhysicalActivity	75.92	10.60
Age	74.19	10.52
Gender (Female)	42.45	10.91
Smoking (Yes)	26.35	7.34
Genetic Risk (Medium)	0.00	0.08

Table 2: Feature importance rankings from random forest (RF) and gradient boosting (GBM) base learners. RF values normalized to 0-100 scale; GBM values are relative importance percentages.

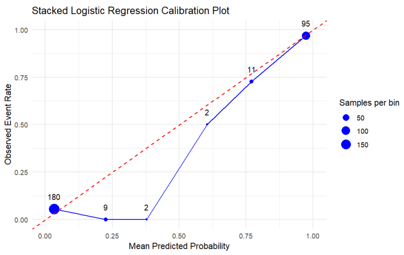


Figure 3: Calibration plot for the stacked model.

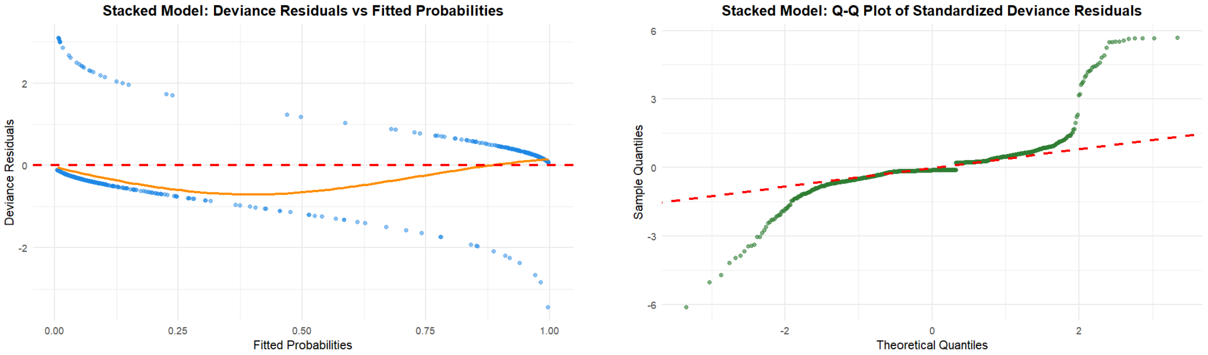
Although regularization in the logistic regression layer can control model complexity to some extent, it is difficult to completely offset the high correlation between the predictions of the base models. Future work could explore more diverse machine learning methods, such as appropriate deep learning models, to provide complementary information, or build more complex stacking structures, such as multi-layered or hierarchical ensemble frameworks.

Feature analysis across both rf and RMB, suggesting that routinely collected variables can provide meaningful cancer risk information. This supports population-level risk screening strategies and provides a framework for future clinical applications.

6 Conclusion

This study systematically compared seven classification models and demonstrated that a stacked logistic regression approach using random forest and gradient boosting as base models can improve predictive performance, particularly recall and F1-score, while maintaining interpretability. The stacked model achieved a balanced trade-off between precision and recall, which is crucial for cancer risk assessment where missing high-risk individuals has severe clinical consequences.

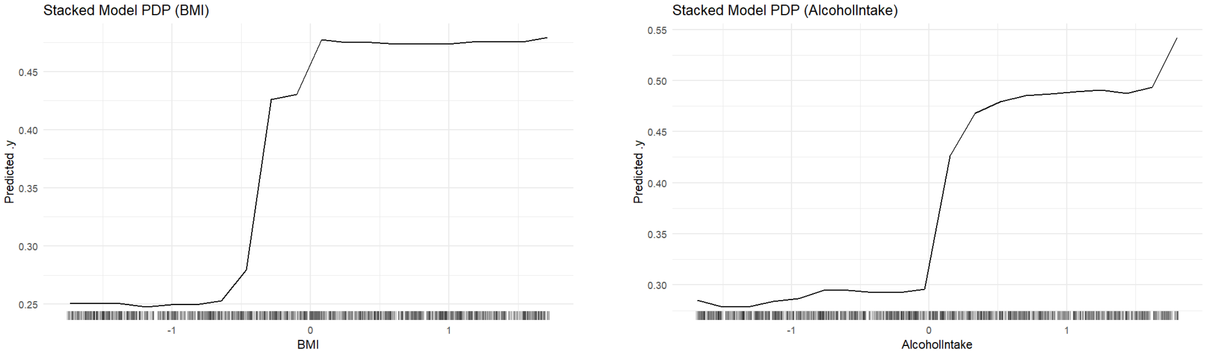
The stacked model achieved high and similar AUC values on both the training and test sets, indicating good generalization ability and no obvious overfitting. In addition, stacked logistic regression showed a significant advantage in probability calibration, reflected in its lower Brier score and smoother calibration curve. By using the prediction results of random forest and gradient boosting models as input, the stacked framework stabilized the probability estimate while preserving the in-



(a) Deviance Residuals vs Fitted Probabilities

(b) Q-Q Plot of Standardized Residuals

Figure 4: Residual diagnostics for the stacked model: (a) deviance residuals vs fitted probabilities, (b) Q-Q plot of standardized residuals.



(a) Partial Dependence Plot: BMI

(b) Partial Dependence Plot: Alcohol Intake

Figure 5: Partial dependence plots (PDP) for BMI and alcohol intake from the stacked model. Note that BMI and Alcohol Intake values are normalized.

terpretability of logistic regression, generating a more reliable risk score. Therefore, this method has higher practical value in cancer risk assessment scenarios that require a balance between discriminative performance, probabilistic stability, and interpretability.

Feature importance analysis across both base learners consistently identified cancer history as the most critical predictor, with lifestyle factors (BMI, alcohol intake, physical activity) and genetic risk also ranking highly, suggesting that routinely collected variables can provide meaningful cancer risk information even without detailed clinical data. These findings support the potential application of stacked machine learning methods for population-level cancer risk screening and provide a framework for future clinical studies exploring personalized risk assessment.

Author Contributions

Q.O. wrote the code and implemented the models. W.S. performed exploratory data analysis and wrote the Data, Discussion, and Conclusion sections. J.Z. wrote all remaining sections of the report.

References

Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 161–168, New York, NY, USA. Association for Computing Machinery.

Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.