Explainable Machine Learning For Personalized Treatment and Lifestyle Recommendations In Cardiovascular Care

THE UNIVERSITY of NEW ORLEANS

Anh Nguyen (annguye9@uno.edu), Md Wasi UI Kabir (mkabir3@uno.edu), Md Tamjidul Hoque (thoque@uno.edu)
Department of Computer Science, University of New Orleans, New Orleans, LA, USA

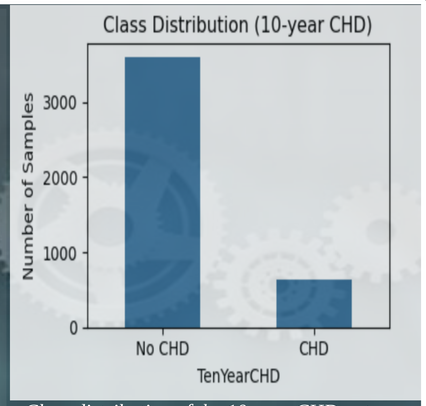
Bioinformatics & Machine Learning Lab

I. INTRODUCTION

- Cardiovascular disease (CVD) remains one of the leading causes of death globally, emphasizing the need for effective predictive and preventive strategies.
- This study develops a machine learning–driven framework for personalized cardiovascular risk modeling and lifestyle guidance using clinical data.
- This work is summarized: a classification based prediction model is implemented to predict the CVD risk, and the generative adversarial network (GAN) is incorporated to learn the underlying relationship between risk factors.

II. DATA SET

- Clinical data obtained from Kaggle, contains 4.3k samples
  - Features: Demographic, clinical, lifestyle
  - Target: 10-year CHD event
- Models were trained and evaluated using 10-fold cross-validation.



Class distribution of the 10-year CHD outcome

III. METHODS

Overall proposed methodology

Raw Data → Prediction models of CVD risks → Lifestyle recommendation with GAIN

Data Preprocessing:

Dataset → KNN Imputation, Scaling/Normalization → Feature Engineering → Feature Selection / PCA → Imbalance handling → Original / PCA → 10-fold Cross-Validation → Training Models

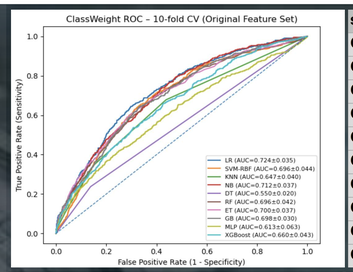
IV. PREDICTIVE MODELS

- **Linear & Distance-based Models**
  - Logistic Regression, SVM, KNN
  - Interpretable baselines for cardiovascular risk prediction
- **Tree-based Ensembles**
  - Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost
  - Capture nonlinear interactions among clinical risk factors
- **Neural Network**
  - Multi-layer Perceptron (MLP)
  - Learns nonlinear feature representations

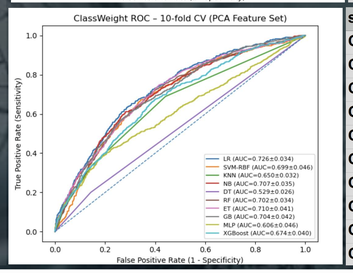
V. EVALUATION AND RESULTS

- **Original Feature Set:** Nonlinear and ensemble-based models outperform linear baselines, highlighting the importance of modeling complex feature interactions.
- **PCA Feature Set:** PCA maintains predictive performance while improving model stability across folds.

***Across all features representations:** Model performance remains consistent across both Original and PCA representations, indicating robust and generalizable learning patterns.



strategy	dataset	model	ROC_AUC (mean±std)	F1 (mean±std)	ACC (mean±std)
ClassWeight	Original	Logistic Regrssion	0.724 ± 0.035	0.378 ± 0.026	0.667 ± 0.017
ClassWeight	Original	Naive Bayes	0.712 ± 0.037	0.318 ± 0.036	0.801 ± 0.012
ClassWeight	Original	Extremely Randomized Trees	0.700 ± 0.037	0.117 ± 0.040	0.852 ± 0.005
ClassWeight	Original	Gradient Boosting	0.698 ± 0.030	0.148 ± 0.043	0.849 ± 0.007
ClassWeight	Original	Support Vector Machine	0.696 ± 0.044	0.371 ± 0.041	0.693 ± 0.022
ClassWeight	Original	Random Forest	0.696 ± 0.042	0.052 ± 0.046	0.849 ± 0.005
ClassWeight	Original	Extreme Gradient Boosting	0.660 ± 0.043	0.313 ± 0.038	0.749 ± 0.018
ClassWeight	Original	K-Nearest Neighbors	0.647 ± 0.040	0.163 ± 0.055	0.833 ± 0.009
ClassWeight	Original	Multilayer Perceptron	0.613 ± 0.063	0.256 ± 0.050	0.801 ± 0.016
ClassWeight	Original	Decision Tree	0.550 ± 0.020	0.236 ± 0.035	0.767 ± 0.019



strategy	dataset	model	ROC_AUC (mean±std)	F1 (mean±std)	ACC (mean±std)
ClassWeight	PCA	Logistic Regrssion	0.726 ± 0.034	0.381 ± 0.023	0.676 ± 0.017
ClassWeight	PCA	Naive Bayes	0.710 ± 0.041	0.061 ± 0.045	0.849 ± 0.005
ClassWeight	PCA	Extremely Randomized Trees	0.707 ± 0.035	0.236 ± 0.051	0.828 ± 0.007
ClassWeight	PCA	Gradient Boosting	0.704 ± 0.042	0.124 ± 0.039	0.845 ± 0.006
ClassWeight	PCA	Support Vector Machine	0.702 ± 0.034	0.018 ± 0.021	0.847 ± 0.003
ClassWeight	PCA	Random Forest	0.699 ± 0.046	0.369 ± 0.040	0.690 ± 0.021
ClassWeight	PCA	Extreme Gradient Boosting	0.674 ± 0.040	0.313 ± 0.042	0.775 ± 0.022
ClassWeight	PCA	K-Nearest Neighbors	0.650 ± 0.032	0.184 ± 0.047	0.835 ± 0.010
ClassWeight	PCA	Multilayer Perceptron	0.606 ± 0.046	0.260 ± 0.025	0.797 ± 0.012
ClassWeight	PCA	Decision Tree	0.529 ± 0.026	0.200 ± 0.045	0.757 ± 0.025

VI. GENERATIVE ADVERSARIAL IMPUTATION NETWORKS

- GAIN learns conditional dependencies among cardiovascular risk factors to enable realistic lifestyle simulations.

GAIN Models → Generation: High Risk Individual → Applicable Lifestyle Modification → Recommended Optimal Lifestyle Modification

- Variable groups in GAIN:

Variable	Features
X^U - Unchangeable Variables	Sex, Age, Education, stroke history
X^D - Directly Changeable Variables	Smoking, Medication, Cholesterol, BMI
X^I — Indirectly Changeable Variables	Blood Pressure, Heart rate

- GAIN performs two tasks:
  - Generate applicable alternative lifestyle behaviors for the high-risk patients.
  - Simulate feasible physiological responses (e.g., blood pressure, heart rate) associated with lifestyle changes for the high-risk patients.

High-Risk Individual → GAIN ($X^U \rightarrow X^D \rightarrow X^I$) → Simulated Lifestyle → Risk Evaluation

VII. CONCLUSION/ FUTURE WORK

- Tree-based ensemble models achieve the strongest and most stable predictive performance for cardiovascular risk prediction across multiple feature representations. These models provide a reliable foundation for personalized risk simulation.
- This is most likely because the dataset is noisy, and a tree-based approach, by taking the average outcome, reduces variance error, and results in a better outcome.
- Future work (a) will complete the full GAIN-based lifestyle recommendation pipeline, including alternative lifestyle generation, indirect physiological variable simulation, and expected utility–based optimization to select personalized lifestyle interventions, (b) collection of a less noisy and larger dataset.

VIII. REFERENCES

- [1] **Dogan, A., Li, Y., Odo, C. P., Sonawane, K., Lin, Y., & Liu, C.** (2023). *A utility-based machine learning-driven personalized lifestyle recommendation for cardiovascular disease prevention.* **Journal of Biomedical Informatics, 141**, 104342. https://doi.org/10.1016/j.jbi.2023.104342
- [2] **Climente-González, H., Oh, M., Chajewska, U., Hosseini, R., Mukherjee, S., Gan, W., Traylor, M., Hu, S., Fatemifar, G., Ghouse, J., Del Villar, P. P., Vernet, E., Koelling, N., Du, L., Abraham, R., Li, C., … Howson, J. M. M.** (2025). *Interpretable machine learning leverages proteomics to improve cardiovascular disease risk prediction and biomarker identification.* **Communications Medicine, 5**, Article 170. https://doi.org/10.1038/s43856-025-00872-0

ACKNOWLEDGE

Research reported in this poster was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424-21.