ASU College of Health Solutions
Arizona State University

EMBEDDED MACHINE INTELLIGENCE LAB

Enhancing Metabolic Syndrome Prediction with Hybrid Data Balancing and Counterfactuals

Sanyam Paresh Shah, Abdullah Mamun, Shovito Barua Soumma, Hassan Ghasemzadeh

Embedded Machine Intelligence Lab

Arizona State University

ghasemzadeh.com

Full-Text PDF

Introduction

❖ *What is Metabolic Syndrome?*

- Cluster of risk factors: obesity, dyslipidemia, hypertension, insulin resistance
- Global prevalence >25% in adults
- Significantly increases CVD and T2DM risk

❖ *Current Challenges*

- Class Imbalance in Datasets
- Data Scarcity and missing values
- Methodological inconsistencies
- Limited interpretability of clinical use

Existing Methods for Addressing Class Imbalance

- ***Strategy I***
  - Models trained on the **original imbalanced dataset**.
  - No oversampling applied.

- ***Strategy II***
  - **Random oversampling** applied **only to training set**.

- ***Strategy III***
  - **Balance data with help of synthetic data**
  - **e.g., SMOTE, ADASYN**
  - **Recent methods based on generative models: BIDC2, AIMEN**

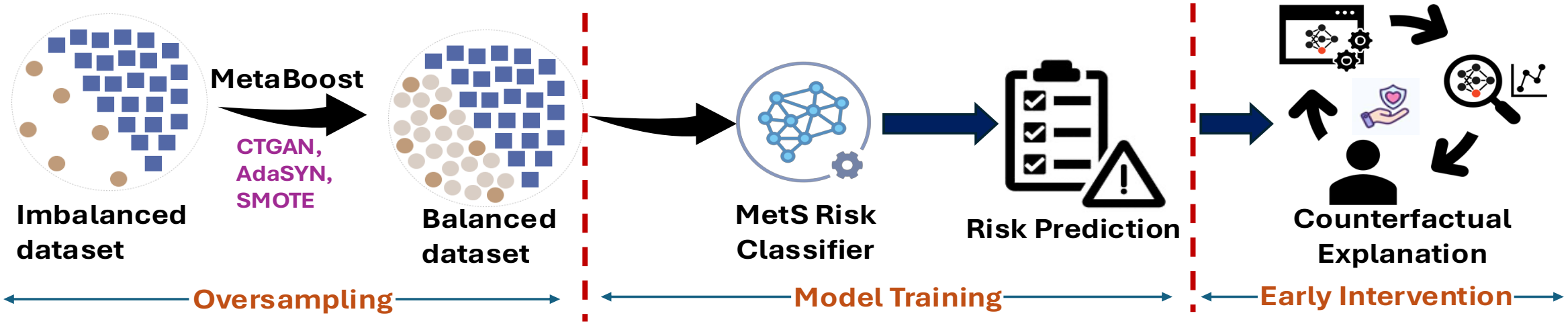
Research Question

- ***How can hybrid ML approaches with advanced data balancing and counterfactual analysis enhance MetS prediction and clinical interpretability?***

Our Proposed System: MetaBoost

❖ ***Advanced Techniques Explored***

SMOTE, **ADASYN**, **CTGAN** used individually and in hybrid forms.



Dataset and Preprocessing

❖ *NHANES Dataset*

- 2,401 individuals with 13 clinical features
- Features: age, sex, waist circumference, BMI, blood glucose, HDL, triglycerides, etc.
- Target: MetS presence/absence

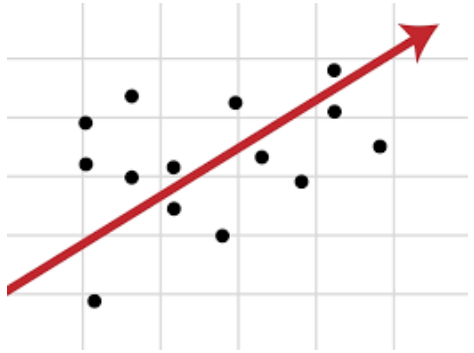
❖ *Preprocessing*

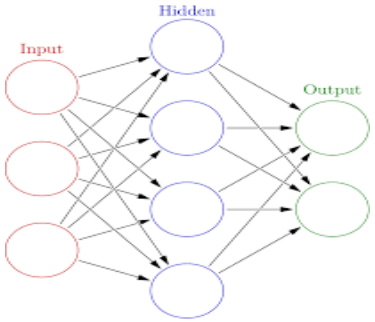
- Removed marital status (8.66% missing values)
- Categorical encoding: Sex (Male=0, Female=1), Race (White=0 to Other=5)
- Mean imputation for Income, WaistCirc, BMI
- 67%/33% train/test split with balanced test set

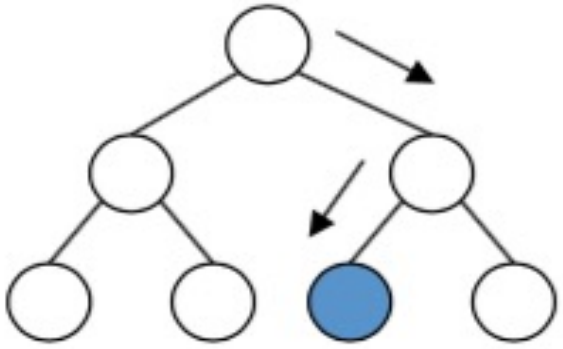
Model Evaluation and Performance

❖ ***Machine Learning Models Tested***

- XGBoost Classifier
- Random Forest
- TabNet Logistic
- Regression
- Multi-Layer Perceptron (MLP)
- Decision Tree







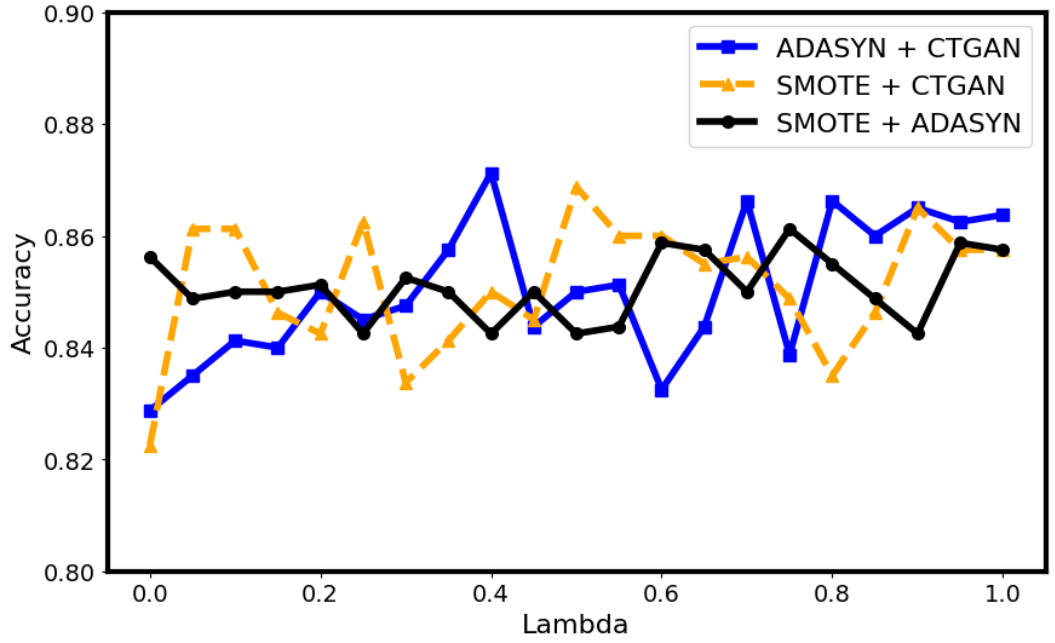
❖ ***Evaluation Metrics***

- Accuracy, Precision, Recall, F1 Score

Performance Comparison

Across Different ROS Strategies

		RF	DT	XGB	LR	MLP	TNet
Without ROS	Acc	0.804	0.815	**0.843**	0.744	0.545	0.673
	Pre	0.931	0.879	**0.936**	0.920	0.914	0.675
	Rec	0.656	0.732	0.736	0.534	0.270	**0.920**
	F1	0.770	0.799	**0.824**	0.676	0.417	0.779
ROS on Training Set	Acc	0.827	0.815	**0.859**	0.798	0.629	0.811
	Pre	**0.917**	0.871	0.913	0.775	0.579	0.769
	Rec	0.719	0.741	0.793	0.838	**0.945**	0.890
	F1	0.806	0.801	**0.849**	0.805	0.718	0.825



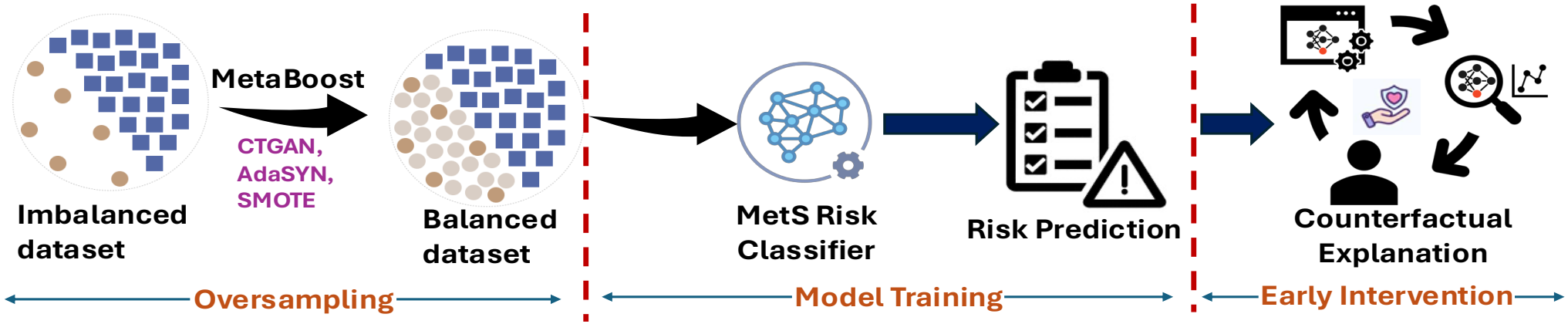
Results of MetaBoost (with XGBoost backbone)

Method	Weights	Accuracy	Precision	Recall	F1
SMOTE	-	0.868	0.889	0.840	0.864
ADASYN	-	0.855	0.872	0.833	0.852
CTGAN	-	0.866	**0.913**	0.810	0.858
ADASYN+CTGAN	(0.4, 0.6)	**0.871**	0.890	**0.848**	**0.868**
SMOTE+CTGAN	(0.5, 0.5)	0.869	0.891	0.840	0.865
SMOTE+ADASYN	(0.75, 0.25)	0.861	0.877	0.840	0.858
SMOTE+CTGAN+ADASYN	(0.05,0.55,0.4)	0.869	0.889	0.843	0.865

MetaBoost

- ***Individual techniques***
  - *SMOTE: Synthetic Minority Oversampling Technique*
  - *ADASYN:* Adaptive Synthetic Sampling (focuses on decision boundary)
  - *CTGAN*: Conditional Tabular Generative Adversarial Networks

- ***Hybrid Approach***
  - Weighted combination of synthetic data from multiple methods
  - Systematic weight optimization (0.05 increments)
  - Two-method combinations: 20 different weight combinations
  - Three-method combination (SMOTE + ADASYN + CTGAN): 235 different weight combinations



Counterfactual Analysis

- Nearest Instance Counterfactual Explanations (NICE) algorithm
- L1 norm for feature-wise distance measurement
- ***Data Analysis***
  - Normalized average distance, standard deviation, average feature changes, and percentage of altered features were computed.
- ***Visualization***
  - A Random Forest Classifier was applied to visualize decision boundaries between original and counterfactual instances.
  - PCA-transformed and standardized data were used for visualization.

Counterfactual Analysis

Key Findings

- Average normalized distance: 1.489 (±1.120)
- Average features modified: 2.054 (±1.070)
- Only 17.1% of features need changes for class flip

Metric	Value
Average Normalized Distance	1.489
Standard Deviation of Normalized Distance	1.120
Average Sparsity	2.054
Standard Deviation of Sparsity	1.070
Percentage of Features Changed	17.1%

Clinical Interpretability

Minimal feature modifications needed for risk category changes

Counterfactual Analysis

❖ *Most Frequently Modified Features*

- **Blood Glucose**: 50.3% (most critical)
- **Triglycerides**: 46.7% (second most important)
- **Waist Circumference**: 42.9%
- **HDL**: 33.7%

❖ *Rarely Modified Features*

- **Demographics**: Sex (0.1%), Race (0%)
- **Medical**: Albuminuria (0.1%)
- **Socioeconomic**: Income (1.7%)

Feature	Change Rate (%)
BloodGlucose	50.3%
Triglycerides	46.7%
WaistCirc	42.9%
HDL	33.7%
BMI	9.6%
Age	8.9%
UrAlbCr	7.8%
UricAcid	3.5%
Income	1.7%
Sex	0.1%
Albuminuria	0.1%
Race	0.0%

Clinical Significance

Model focuses on **modifiable metabolic factors** rather than fixed demographic characteristics

PCA Analysis

❖ *PCA-Reduced Space Analysis*

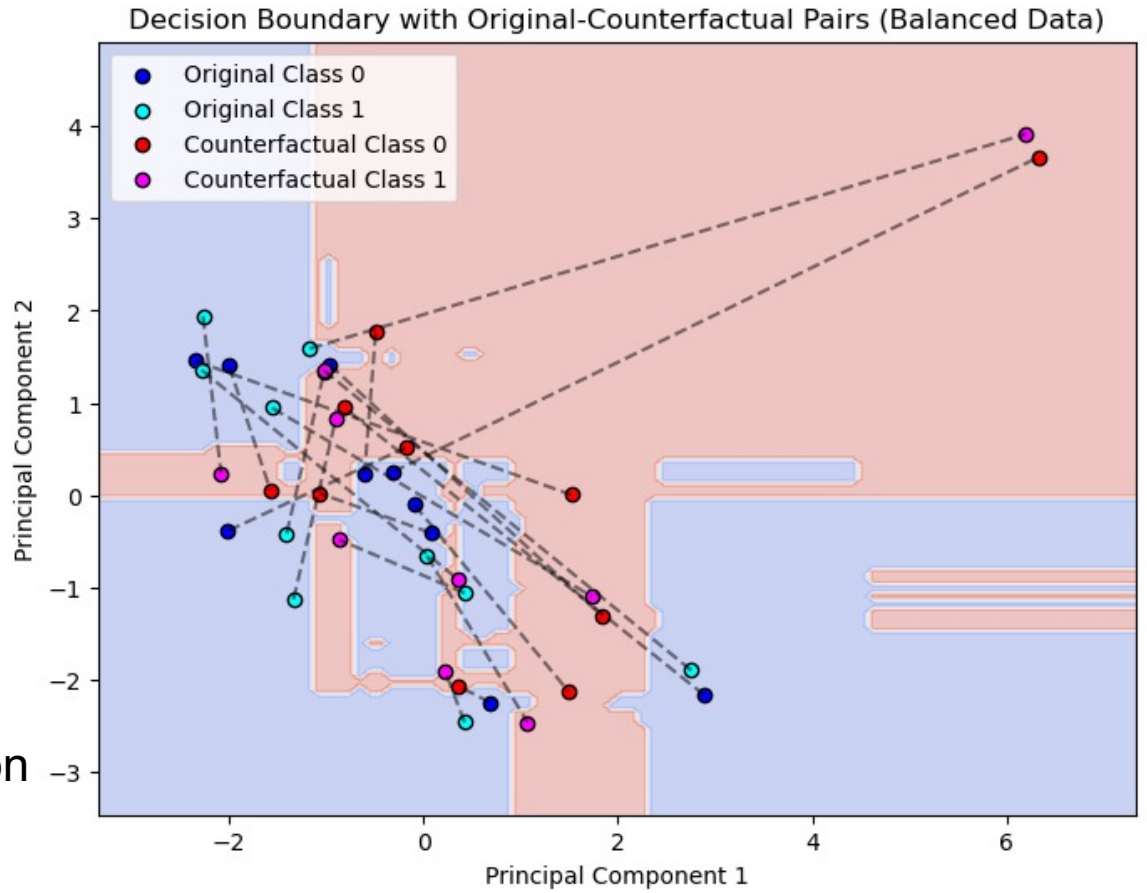
- Original instances clustered in central region (-2 to 2)
- Counterfactual instances show wider dispersion
- Complex, non-linear decision boundaries revealed

❖ *Random Forest Classifier Patterns*

- Multiple disjoint decision regions
- Variable transition lengths between classes
- Local pattern capture capability demonstrated

❖ *Clinical Translation*

Different patients require different degrees of intervention based on their position in feature space



Summary

❖ *Key Contributions*

- **MetaBoost framework**: Novel hybrid data balancing approach (1.87% accuracy improvement over individual methods)
- **Performance achievement**: 87.1% accuracy, 0.868 F1-score
- **Clinical interpretability**: Counterfactual analysis for actionable insights
- **Evidence-based targeting**: Blood glucose and triglycerides as primary intervention points

❖ *Clinical Significance*

- Addresses critical healthcare challenges: class imbalance, data scarcity
- Provides interpretable ML models for clinical decision-making
- Enables personalized intervention strategies

❖ *Impact*

- Advances methodological rigor in MetS prediction while providing actionable clinical insights for mitigating global metabolic syndrome burden

College of Health Solutions
Arizona State University

EMBEDDED MACHINE INTELLIGENCE LAB

Thank You!




Full-Text PDF

https://ghasemzadeh.com/

hassan.ghasemzadeh@asu.edu