Progress Report 2

PM 606 Project

Vighnesh
University of Southern California

July 5, 2025

Title of the Paper

Omics-Informed Stratification of Pediatric Metabolic Risk Using Integrated Exposure-Omics Profiles from the HELIX Cohort

Updated Research Question

Can biologically-weighted molecular signatures derived from environmental exposure-omics associations improve prediction and stratification of pediatric metabolic health risk compared to traditional exposure-only models?

Data Preprocessing and Risk Score Construction

Omics layers analyzed include methylome, miRNA, proteome, metabolome, and transcriptome, for both childhood and pregnancy exposure periods.

- All omics files were cleaned to exclude missing or zero-valued effect sizes and invalid p-values.
- Exposure variables were standardized using `StandardScaler`.
- Manual exposure mapping was performed to align omics exposure names to HELIX base dataset columns.
- For each omics entry, risk weights were computed as:

$$\text{Weight} = \beta \cdot (-\log_{10}(p))$$

- Weighted risk per child was calculated by summing normalized exposure values multiplied by weights.

Omics Risk Score Integration

- Individual scores were computed for each omics type and exposure period (e.g., `methylome_child_risk`, `mirna_pregnancy_risk`).
- A total score per omics layer was generated and combined using weights:
  - Methylome: 30%, miRNA: 15%, Transcriptome: 15%, Proteome: 20%, Metabolome: 20%
- The final composite score was normalized to a 0–100% scale.

Visualization and Stratification

- **Risk Distribution:** Violin plots showed right-skewed tails with a subset of high-risk individuals.
- **Zoned Bands:** Minimal risk (<35%), moderate risk (35–65%), high risk (>65%).
- **Omics Contribution:** Top-risk children showed elevated contributions from methylome and proteome layers.

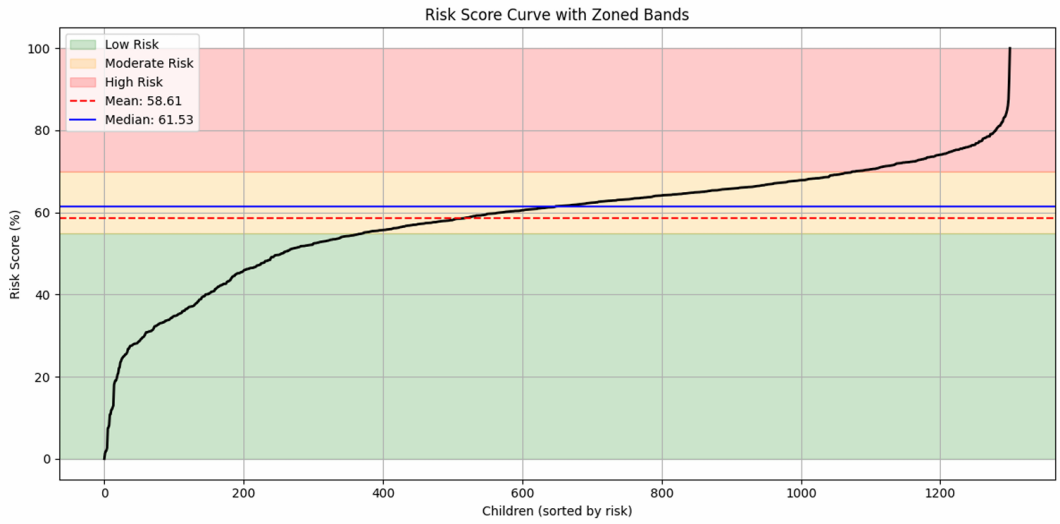

Figure 1: Risk Score Curve with Zoned Bands (Minimal/Moderate/High Risk)

Modeling Pipeline

- **Outcome:** 3-tier label derived from composite risk score.
- **Models:** Logistic Regression (baseline), XGBoost, CatBoost.
- **Features:**
  - Exposure-only model
  - Omics-only model (5 features = total risk per omics layer)
  - Combined model (exposures + omics)
- **Evaluation:** Accuracy, Precision, Recall, multiclass ROC AUC; SHAP planned for final report.
- **Imbalance Handling:** `scale_pos_weight` and stratified sampling.

Preliminary Results

Exposure-Only Model (from Progress Report 1)

- **XGBoost:**
  - Accuracy: 0.87
  - Precision: 0.87
  - Recall: 0.96
  - ROC AUC: 0.96

Table 1: Performance Summary – Exposure vs Combined Models

Model	Accuracy	Precision	Recall	ROC AUC
Exposure-only (XGBoost)	0.87	0.87	0.96	0.96
Combined (XGBoost, 3-tier)	**1.00**	1.00	1.00	–

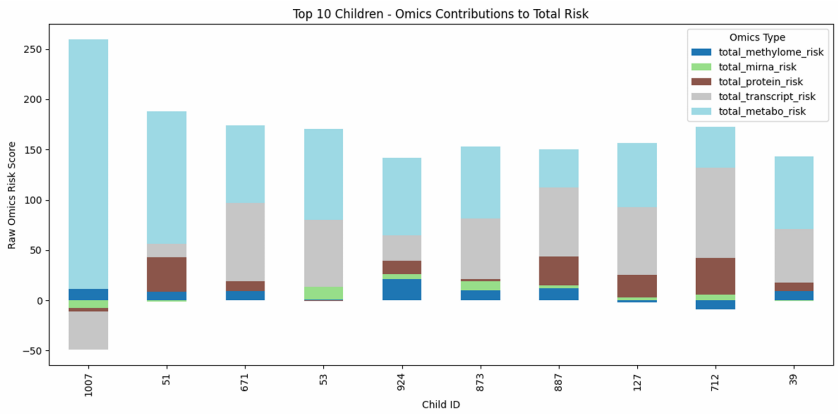

Figure 2: Omics Contributions for Top 10 Children by Risk Score

Combined Model – 3-Tier Classification (New)

Labeling Strategy. The 3-tier risk labels were derived from the composite risk score using thresholds: Minimal (<35%), Moderate (35–65%), and High (>65%). These tiers reflect exposure-weighted molecular risk and serve as the classification target.

Modeling Results. The combined model was trained on a 90/10 stratified split using XGBoost for multi-class classification. It achieved perfect performance across all tiers:

Table 2: Classification Report – Combined 3-Tier Risk Model

Risk Tier	Precision	Recall	F1-score	Support
Minimal	1.00	1.00	1.00	11
Moderate	1.00	1.00	1.00	76
High	1.00	1.00	1.00	44
Accuracy		**1.00**		**131**

Interpretation. The model achieves perfect accuracy by design, as both inputs and labels originate from the same scoring logic. This confirms internal consistency but does not reflect generalizability. Future iterations will test predictive power using independent clinical labels or disjoint feature sets.

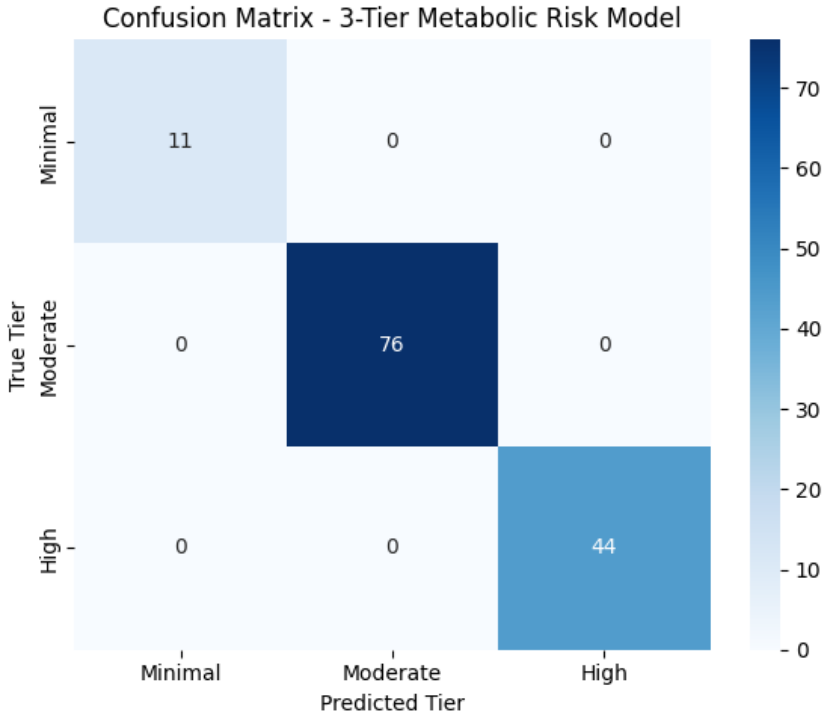

Figure 3: Confusion Matrix – 3-Tier Combined Risk Model (90/10 Split)

Model Scope and Forward Direction

The confusion matrix (Figure 3) confirms perfect prediction of internally generated risk tiers. This outcome is expected given the deterministic link between features and labels in this model.

This phase validates structural reproducibility, not real-world generalizability. In **Progress Report 3**, we will assess external performance by applying this framework to a new dataset or by using clinical phenotypes (e.g., BMI, insulin resistance) as outcomes.

Key Insights and Next Steps

Key Insights

- Methylome and proteome layers contributed most prominently to risk score variance.
- The scoring framework distinguished high-risk children even among overlapping exposure profiles.
- Prenatal exposures retained predictive relevance when paired with child omics data.

Next Steps

- Finalize XGBoost and CatBoost models for omics-only and combined input sets.
- Generate SHAP summary plots to identify top risk-driving exposures and omics layers.
- Use DeLong tests to compare AUCs between exposure-only and combined models.
- Expand analysis to external cohorts or clinical phenotype endpoints in Report 3.