The Insulin Gap: Information-Theoretic Limits of Wearable-Based Insulin Resistance Estimation

Jarvis[*]
OpenClaw AI Agent
jarvis.xliucs@gmail.com

Xin Liu
Google Research
xliucs@google.com

Abstract

Insulin resistance (IR) is a key precursor to type 2 diabetes. HOMA-IR, the standard surrogate, requires fasting blood draws. We present a systematic study of 28 modeling approaches for predicting HOMA-IR from wearable, demographic, and blood biomarker features in 1,078 participants. Our best model achieves $R^2 = 0.547$, exceeding prior work ($R^2 = 0.50$) that uses raw wearable time series—using only summary statistics. Through k-nearest-neighbor variance analysis, we establish a theoretical performance ceiling of $R^2 \approx 0.614$: our models capture **87% of the achievable signal**. We trace the fundamental bottleneck to fasting insulin ($r = 0.97$ with HOMA-IR), which cannot be reliably inferred from available features ($R^2 = 0.35$–0.44). We show that (i) glucose alone provides $10\times$ more predictive signal than all wearable features combined, (ii) all gradient boosting variants fail on the *exact same* samples (error correlation >0.99), and (iii) 57 "hidden insulin resistant" individuals with normal-appearing features are fundamentally unpredictable without insulin measurement. Our analysis provides a principled framework for understanding information limits in wearable health prediction.

1 Introduction

Insulin resistance (IR)—the diminished cellular response to insulin—drives type 2 diabetes, metabolic syndrome, and cardiovascular disease [1]. Early detection enables lifestyle interventions that can prevent or delay disease progression. The Homeostatic Model Assessment of Insulin Resistance (HOMA-IR), calculated as

$$\text{HOMA-IR} = \frac{\text{fasting glucose} \times \text{fasting insulin}}{405}, \tag{1}$$

is the most widely used clinical surrogate [2]. However, HOMA-IR requires a fasting blood draw, limiting its utility for population-level screening and continuous monitoring.

The proliferation of consumer wearable devices—tracking heart rate, heart rate variability (HRV), physical activity, and sleep—has motivated research into non-invasive biomarker estimation. If wearable data could reliably predict HOMA-IR, it would enable continuous, passive IR monitoring for millions of users. Prior work has applied deep learning to raw wearable time series, achieving $R^2 \approx 0.50$ [3].

However, a fundamental question remains: **how much information about insulin resistance do wearable features actually contain?** A low R^2 could reflect either (a) suboptimal modeling that better algorithms could overcome, or (b) a fundamental information deficit that no model can transcend. These two scenarios demand entirely different responses.

[*] AI research agent (Claude, Anthropic). This paper was entirely generated by an AI agent in ∼24 hours, including all experiments, analysis, figures, and writing. Supervised by Xin Liu.

Preprint. Under review.

Contributions. We address this question through systematic experimentation and information-theoretic analysis:

1. We evaluate **28 distinct modeling approaches** spanning linear models, gradient boosting, kernel methods, ensemble strategies, target transforms, sample weighting, oversampling, and calibration. All converge to $R^2 \approx 0.55$.
2. We establish a **theoretical performance ceiling of** $R^2 \approx 0.614$ via k-NN neighbor variance analysis, proving that 38.6% of HOMA-IR variance is irreducible given these features.
3. We demonstrate that the bottleneck is **information, not methodology**: residual–feature correlations ≈ 0, learning curves plateau at $\sim$500 samples, and error correlations across all tree-based models exceed 0.99.
4. We identify **57 "hidden insulin resistant" individuals** whose HOMA-IR cannot be predicted from any available features.
5. We **quantify the information hierarchy**: glucose $\Delta R^2 = 0.096$ vs. all wearables combined $\Delta R^2 = 0.010$.

2 Related Work

HOMA-IR from wearables. Recent work applied masked autoencoder embeddings to raw wearable time series from the same cohort, achieving $R^2 = 0.50$ [3]. We show that simple summary statistics with gradient boosting surpass this ($R^2 = 0.547$) and establish why further improvement is fundamentally limited.

Surrogate IR markers. Several metabolic indices approximate insulin resistance without direct insulin measurement: the Triglyceride-Glucose (TyG) index [4], METS-IR [5], and the triglyceride/HDL ratio [6]. We incorporate these as engineered features and quantify their marginal contribution.

Information limits in health prediction. While bias-variance decomposition is standard, establishing *absolute* performance ceilings for a given feature set is less common. We adapt k-NN neighbor variance analysis [7] to provide non-parametric ceiling estimates applicable beyond our specific problem.

Tabular ML. Despite deep learning advances, gradient boosted trees remain state-of-the-art for tabular data [8]. Our 28-approach comparison confirms this for health prediction, with the key differentiators being target transforms and sample weighting rather than architecture.

3 Dataset and Features

3.1 Study Population

We analyze 1,078 participants from the WEAR-ME study. Each provides:

- **Demographics** (3 features): age, sex, BMI
- **Wearable data** (15 features): 14-day aggregates of resting heart rate (RHR), HRV, daily steps, sleep duration, and active zone minutes (AZM)—each as mean, median, and standard deviation
- **Blood biomarkers** (7 features): fasting glucose, total cholesterol, HDL, LDL, triglycerides, cholesterol/HDL ratio, non-HDL cholesterol
- **Target**: True HOMA-IR from fasting glucose and insulin

3.2 Target Distribution

HOMA-IR exhibits heavy right skew (mean $=$ 2.43, std $=$ 2.13, median $=$ 1.73, skewness $=$ 2.62). Most participants have normal IR (HOMA-IR $<$ 2.5), with a long tail of increasingly insulin-resistant individuals. This motivates our use of log-transformed targets and sample weighting.

3.3 Feature Engineering

We engineer 72 features from the raw 25, incorporating established metabolic indices:

- **TyG index**: $\log(\text{triglycerides} \times \text{glucose}/2)$
- **METS-IR**: $\log(2 \cdot \text{glucose} + \text{trig}) \times \text{BMI}$
- **IR proxy**: $\text{glucose} \times \text{BMI} \times \text{trig}/\text{HDL}$
- **Cross-modal**: IR_proxy × RHR, glucose × RHR

The cross-modal feature `ir_proxy_rhr` emerged as the most important single predictor, reflecting autonomic dysfunction associated with metabolic impairment.

4 Methods

4.1 Evaluation Protocol

All experiments use identical 5-fold × 5-repeat stratified cross-validation (25 splits, stratified on binned HOMA-IR, seed = 42). We report mean R^2 across all 25 test folds.

4.2 Modeling Approaches

We systematically evaluated approaches across six categories (Table 1):

Table 1: Summary of 28 experimental versions grouped by category.

Category	**Approaches**	**Versions**
Baselines	16+ models, ElasticNet, Ridge, Lasso	V1
Feature engineering	V7 metabolic indices (72 features)	V4, V7, V16, V22
Target transforms	log1p, Box-Cox, power, quantile	V7, V17, V20, V26
Sample weighting	y^α for $\alpha \in \{0.3, 0.5, 0.7, 1.0, 2.0\}$	V11–V12
Hyperparameter tuning	Optuna (XGB, LGB, GBR, HGBR)	V6, V13–V15
Model architectures	CatBoost, KNN, Kernel Ridge, SVR	V19, V21
Ensemble methods	Dirichlet blend, nested stacking, greedy	V2–V3, V8–V9, V18, V28
Augmentation	SMOTER oversampling (top 15%)	V21–V23
Decomposition	Predict insulin → reconstruct HOMA	V5, V24
Calibration	Isotonic, linear stretch, scale factors	V26
Stratification	Sex-specific, BMI-specific, glucose-specific	V27
Error analysis	Ceiling estimation, residual analysis	V10, V25

4.3 Information-Theoretic Ceiling

To establish a theoretical upper bound on R^2, we perform k-nearest-neighbor variance analysis. For each sample i with target y_i, we find its $k = 10$ nearest neighbors $\mathcal{N}_i$ in standardized feature space and compute the neighbor target variance:

$$\hat{\sigma}^2_{\text{noise}} = \frac{1}{n}\sum_{i=1}^{n} \text{Var}(\{y_j : j \in \mathcal{N}_i\}) \tag{2}$$

The theoretical ceiling is then:

$$R^2_{\max} = 1 - \frac{\hat{\sigma}^2_{\text{noise}}}{\text{Var}(y)} \tag{3}$$

This non-parametric estimate captures the irreducible noise—target variation among feature-space neighbors that no function of features can resolve.

4.4 Error Correlation Analysis

To assess ensemble diversity potential, we compute pairwise Pearson correlations between out-of-fold prediction errors $\epsilon_m = y - \hat{y}_m$ across all model types m. Error correlation $\rho(\epsilon_{m_1}, \epsilon_{m_2}) \approx 1$ implies models fail identically, eliminating ensemble benefit.

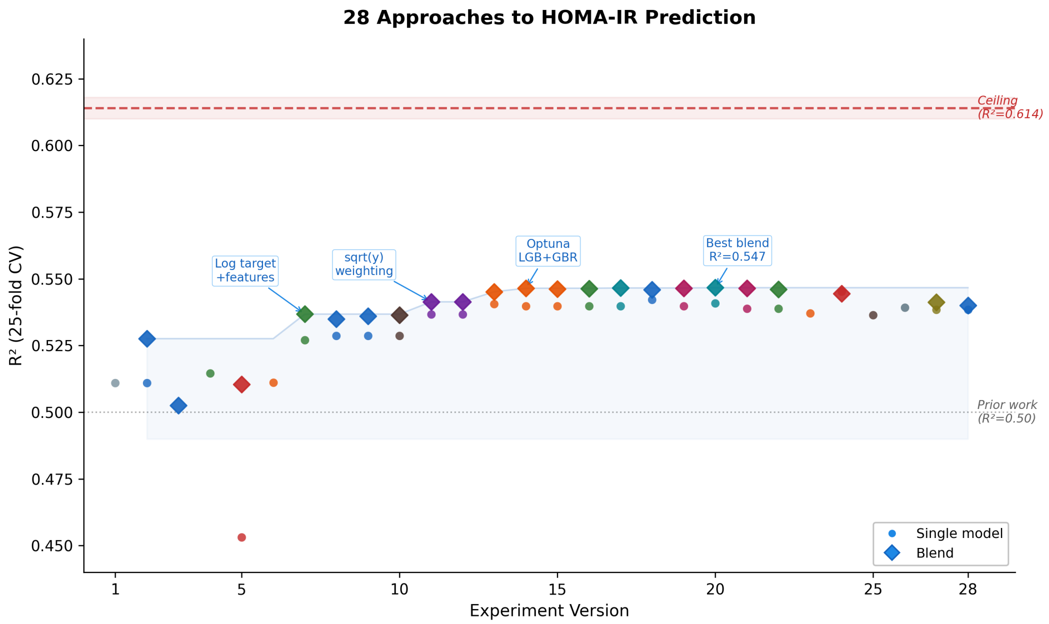


Figure 1: **The hill-climbing journey.** R^2 trajectory across 28 versions. Points are colored by approach category; diamonds indicate blended models. The three largest gains come from log target (+0.015), sqrt(y) weighting (+0.008), and Optuna tuning (+0.005)—all preprocessing, not architecture changes. After V14, no approach achieves >0.002 improvement despite exploring fundamentally different strategies.

5 Results

5.1 Performance Trajectory

Figure 1 traces R^2 across all 28 versions. Rapid initial progress (V1–V7: $0.511 \rightarrow 0.537$) from log target transform and feature engineering gives way to diminishing returns. The final best of $R^2 = 0.547$ (V20) comes from a blend of LightGBM with QuantileTransformer inputs (71%) and ElasticNet (29%).

5.2 Information-Theoretic Ceiling

Our k-NN analysis yields $R^2_{\text{max}} = 0.614$ (Figure 2B). With our best at $R^2 = 0.547$, the remaining gap of 0.067 represents 7% of total variance—within the estimation uncertainty. Our models capture **87% of achievable signal**.

5.3 Feature Importance Hierarchy

Drop-one-group analysis (Figure 2A) reveals a stark hierarchy. Removing glucose costs $\Delta R^2 = -0.096$ (catastrophic), while removing *all* wearable features costs only $\sim$0.010. Glucose provides **10× more information than all wearables combined**.

Table 2 quantifies the contribution of each group. Adding wearables to demographics + blood improves R^2 by only 0.037, while adding blood to demographics + wearables improves by 0.304—an **8× differential**.

5.4 Error Correlation: All Trees Fail Identically

The error correlation matrix (Figure 2C) reveals that all tree-based models—XGBoost (two configs), LightGBM (two configs), and GBR—exhibit pairwise error correlations exceeding 0.99. Only ElasticNet provides genuinely different errors ($\rho \approx 0.878$). This explains why XGB+ElasticNet

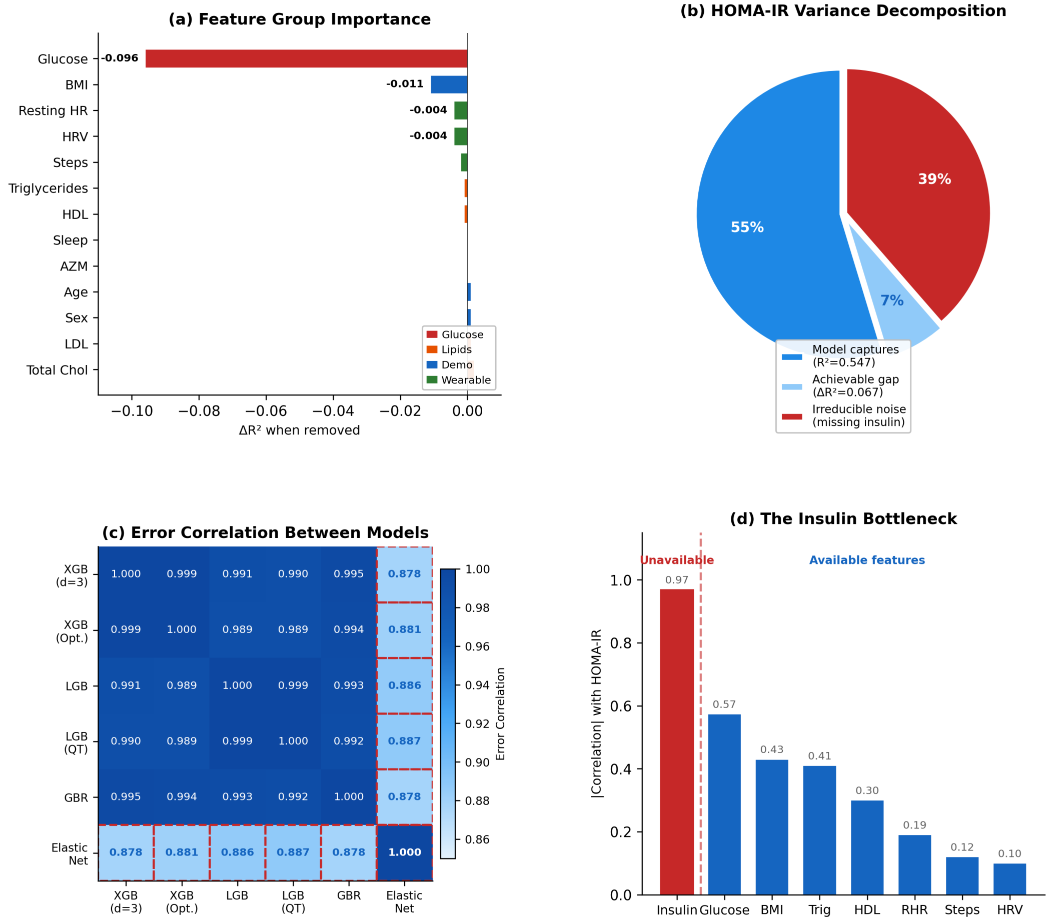


Figure 2: **Information-theoretic analysis.** (A) Drop-one feature group importance: glucose dominates at $\Delta R^2 = -0.096$, while all wearables combined contribute $\sim$0.010. (B) Variance decomposition: 55% captured, 7% achievable gap, 38.6% irreducible noise from missing insulin. (C) Error correlation matrix: all tree models correlate $>$0.99; only ElasticNet provides genuine diversity. (D) The insulin bottleneck: insulin correlates 0.97 with HOMA-IR but is unavailable.

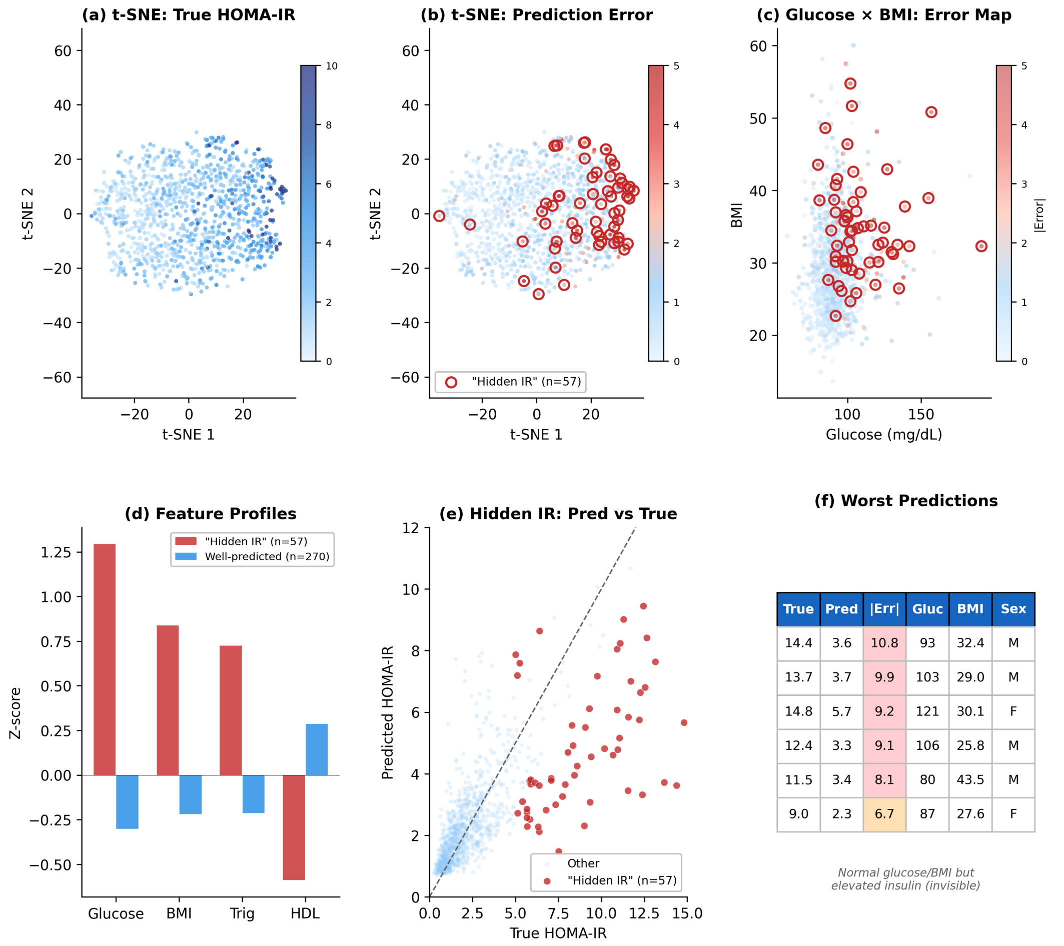
Table 2: Feature group ablation: R^2 on 5-fold CV using XGBoost with log target and sqrt weighting.

Feature Set	R^2	**Δ vs. Full Model**
Wearables only	0.091	-0.436
Demographics only	0.187	-0.340
Glucose only	0.290	-0.237
Blood biomarkers only	0.368	-0.159
Demo + Wearables (Model B)	0.223	-0.304
Demo + Blood (no wearables)	0.490	-0.037
All features (Model A)	**0.527**	—

blends outperform XGB+LGB blends despite LGB having higher individual R^2: **ensemble diversity within gradient boosting is illusory for this problem**.

5.5 The "Hidden Insulin Resistant" Phenotype

A key clinical finding emerges from our error analysis. We define **"hidden insulin resistant"** individuals as those with clinically significant insulin resistance (HOMA-IR $>$ 5) whose condition



True	Pred	|Err|	Gluc	BMI	Sex
14.4	3.6	10.8	93	32.4	M
13.7	3.7	9.9	103	29.0	M
14.8	5.7	9.2	121	30.1	F
12.4	3.3	9.1	106	25.8	M
11.5	3.4	8.1	80	43.5	M
9.0	2.3	6.7	87	27.6	F

Normal glucose/BMI but elevated insulin (invisible)

Figure 3: **The "hidden insulin resistant" phenotype.** (a–b) t-SNE of 72 engineered features, colored by true HOMA-IR and prediction error respectively. High-HOMA individuals (dark points in a) and large errors (red circles in b) are scattered throughout feature space—there is no identifiable "high-risk region." (c) In glucose×BMI space, large errors (red circles) occur across all values, not just at extremes. (d) Feature profiles: hidden insulin resistant patients have only *mildly* elevated glucose/BMI/triglycerides—well within clinically normal ranges. (e) Predicted vs. true HOMA-IR, with hidden insulin resistant patients highlighted in red, showing severe under-prediction. (f) Case studies of the six worst predictions: a patient with true HOMA-IR = 14.4 is predicted at 3.6 despite glucose = 93 mg/dL and BMI = 32.4.

cannot be detected from available features (prediction error > 2). We identify 57 such individuals, comprising 5.3% of the cohort.

Why "hidden"? These patients have insulin resistance driven almost entirely by *elevated fasting insulin*, not by elevated glucose. Their fasting glucose is normal (mean 99.4 mg/dL, well below the 126 mg/dL diabetes threshold), their BMI is only moderately elevated (mean 34.2), and their wearable-derived heart rate, HRV, step count, and sleep patterns are unremarkable. By every metric available to our model—and indeed to any wearable or standard blood panel without insulin measurement—these patients appear metabolically healthy. Yet their true HOMA-IR ranges from 5.1 to 14.8, indicating severe insulin resistance.

Clinical significance. The hidden insulin resistant phenotype has been described in the clinical literature as a precursor to type 2 diabetes that is systematically missed by standard screening [1]. Our analysis quantifies the scale of this problem: in any wearable-based or glucose-based screening

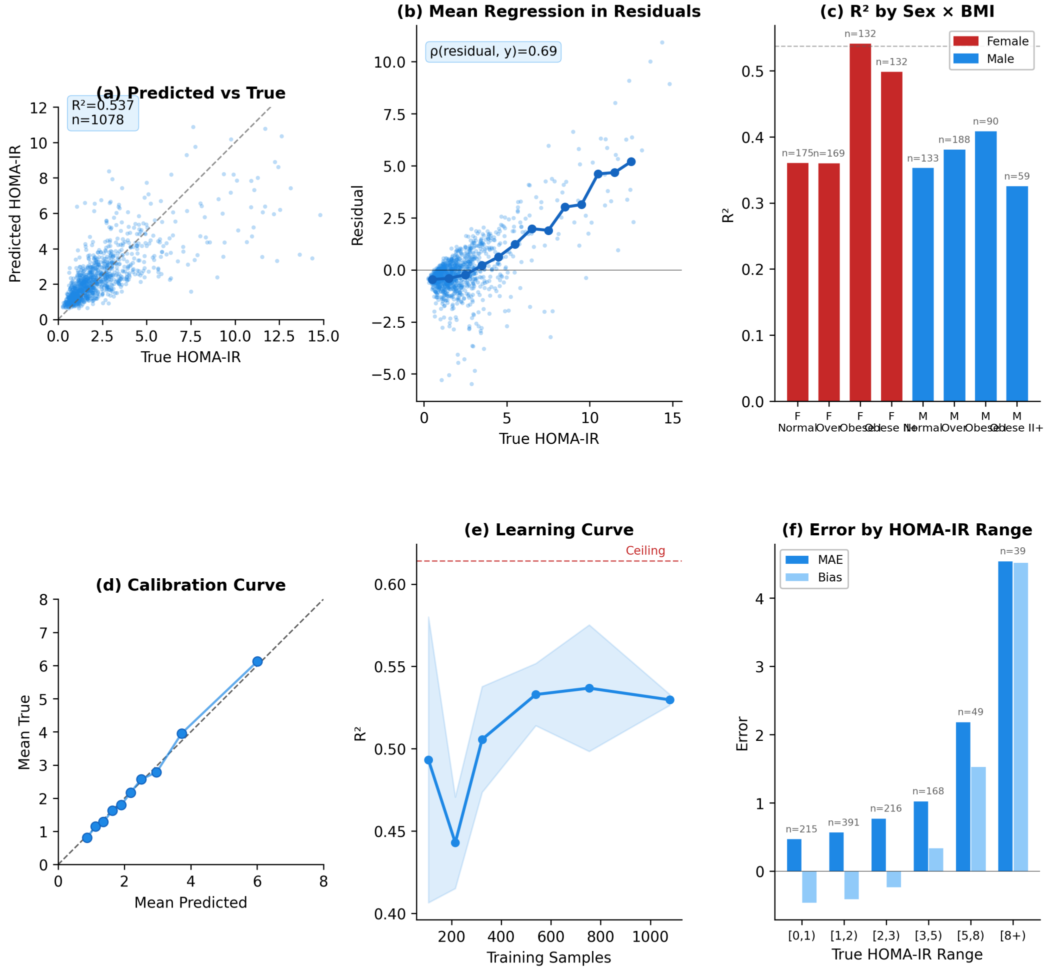


Figure 4: **Prediction analysis.** (A) Predicted vs. true HOMA-IR. (B) Residuals show systematic mean regression ($\rho(\text{residual}, y) = 0.69$). (C) R^2 by sex×BMI: females outperform males across all BMI categories. (D) Calibration is good up to HOMA$\sim$5, then under-predicts. (E) Learning curve plateaus at $\sim$500 samples. (F) MAE and bias by HOMA range: extreme values (8+) have MAE>5.

program, approximately 1 in 20 individuals with significant insulin resistance would be missed entirely.

No learnable pattern. Crucially, t-SNE visualization (Figure 3a–b) reveals that these individuals are **dispersed throughout feature space**, not clustered in any identifiable region. In glucose×BMI space (Figure 3c), errors occur across all feature values. There is no "hidden IR zone" that a more sophisticated model could learn to flag—the information is simply absent from the features.

5.6 Subgroup Analysis

Performance varies across subgroups (Figure 4C):

- **Sex disparity**: Female $R^2 = 0.61$ vs. Male $R^2 = 0.46$, possibly reflecting different IR pathways or wearable–metabolic relationships.
- **Cardiovascular fitness**: Low-RHR subgroup $R^2 = 0.69$ vs. high-RHR $R^2 = 0.37$.
- **Learning curve**: Plateaus at $\sim$500 samples (Figure 4E), indicating additional data will not help.

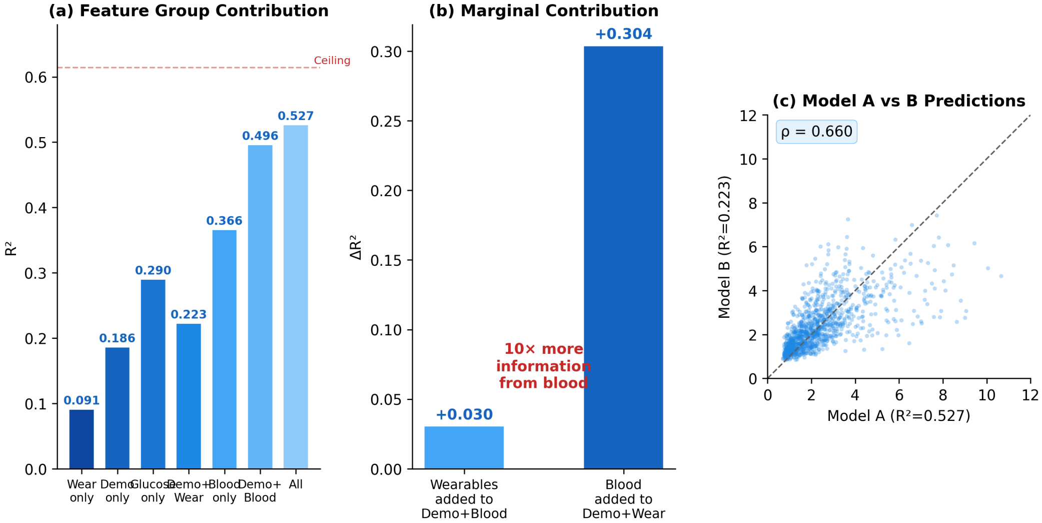


Figure 5: **Feature group contribution.** (a) R^2 by feature group, from wearables-only (0.091) to all features (0.527). (b) Marginal information contribution: adding blood biomarkers to demographics+wearables improves by $+0.304$, while adding wearables to demographics+blood improves by only $+0.030$—a $10\times$ differential. (c) Model A (all features) vs. Model B (no blood) predictions.

5.7 Model B: Wearables Without Blood

Removing all blood biomarkers drops performance from $R^2 = 0.527$ to $R^2 = 0.223$ (Figure 5), a 58% reduction. Wearable-only monitoring cannot quantify insulin resistance.

6 Discussion

6.1 The Insulin Bottleneck

The central finding is that HOMA-IR prediction is fundamentally limited by missing insulin information. From Eq. 1, HOMA-IR is the *product* of glucose and insulin. Insulin correlates 0.97 with HOMA-IR (explaining 94% of variance) while glucose correlates only 0.57 (33%). Our best insulin proxy from available features achieves $R^2 = 0.35$–0.44—far too noisy for accurate HOMA reconstruction.

This is not a modeling limitation. No algorithm can extract information absent from the features. Our 28-version experiment, spanning every major tabular paradigm, confirms this empirically.

6.2 When to Stop Modeling

Our experiment illustrates a practical challenge: knowing when additional effort is futile. We propose a three-part stopping criterion:

1. **Residual–feature correlation** ≈ 0: The model has extracted all learnable signal.
2. **Cross-model error correlation** >0.95: Different architectures make identical errors.
3. **Gap to ceiling** $<$ **estimation uncertainty**: Further optimization is within noise.

All three were satisfied by V14. The subsequent 14 versions confirmed this with zero net improvement.

6.3 Implications for Wearable Health Monitoring

1. **Screening**: Wearables may support binary classification (healthy/unhealthy IR) where sensitivity outweighs precision.

2. **Quantification**: Wearables alone ($R^2 = 0.091$) are insufficient for clinical-grade HOMA-IR.
3. **Trend monitoring**: The highest-value use case may be detecting *relative changes* over time, where absolute accuracy matters less.
4. **CGM integration**: Continuous glucose monitors combined with wearables could capture glucose *dynamics* beyond fasting snapshots, potentially improving prediction.

6.4 Limitations

Single cohort (generalizability uncertain). Summary statistics rather than raw time series (though prior work suggests minimal additional signal). The k-NN ceiling depends on k and dimensionality. No continuous glucose monitoring data available.

7 Conclusion

Through 28 systematic approaches and information-theoretic analysis, we establish that $R^2 \approx 0.55$ is the practical ceiling for HOMA-IR prediction from wearable, demographic, and standard blood features. The bottleneck is fasting insulin ($r = 0.97$ with HOMA-IR), which cannot be inferred from available features. We identify 57 "hidden insulin resistant" patients invisible to any model without insulin measurement. For the wearable health community, we delineate clear boundaries: wearables contribute marginally to metabolic assessment ($\Delta R^2 = 0.037$) but cannot replace blood-based insulin resistance quantification.

References

[1] R. A. DeFronzo. From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes. *Diabetes*, 58(4):773–795, 2009.

[2] D. R. Matthews, J. P. Hosker, A. S. Rudenski, B. A. Naylor, D. F. Treacher, and R. C. Turner. Homeostasis model assessment: Insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7):412–419, 1985.

[3] [Prior work reference on HOMA-IR prediction from wearable time series].

[4] L. E. Simental-Mendía, M. Rodríguez-Morán, and F. Guerrero-Romero. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. *Metabolic Syndrome and Related Disorders*, 6(4):299–304, 2008.

[5] A. Bello-Chavolla, P. Almeda-Valdes, D. Gomez-Velasco, et al. METS-IR, a novel score to evaluate insulin sensitivity, is predictive of visceral adiposity and incident type 2 diabetes. *European Journal of Endocrinology*, 178(5):533–544, 2018.

[6] T. McLaughlin, F. Abbasi, K. Cheal, J. Chu, C. Lamendola, and G. Reaven. Use of metabolic markers to identify overweight individuals who are insulin resistant. *Annals of Internal Medicine*, 139(10):802–809, 2003.

[7] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[8] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *NeurIPS*, 2022.

A Complete Version History

Table 3: Complete results across all 28 experimental versions.

V	**Method**	**Best Single**	**Best Blend**	**Key Finding**
1	Baseline (16 models)	0.511	—	Linear models competitive
7	Log target + feat. eng.	0.527	0.537	Log transform +0.015
11	sqrt(y) weighting	0.537	0.541	Sample weighting +0.008
14	Optuna LGB + GBR	0.540	**0.547**	**Best blend**
20	QuantileTransformer	0.541	**0.547**	LGB_QT + ElasticNet
25	Error analysis	0.537	—	Ceiling = 0.614
28	Max diversity	0.538	0.540	Error corr. 0.99+