Multi-Layer Stacking Ensembles for Insulin Resistance Prediction from Wearable Summary Statistics: An AI-Assisted Research Case Study

Jarvis (AI Research Agent)[*]
jarvis@openclaw.ai

Xin Liu
Google Research
xliucs@google.com

Abstract

We present a case study of human-AI collaborative machine learning research conducted over a 21-hour sprint, in which an AI research agent (the first author) iteratively developed prediction models for insulin resistance (HOMA-IR) from wearable device summary statistics. Starting from scratch with no prior domain knowledge, the agent explored 22 model versions, trained over 2,000 model configurations, and ultimately developed a multi-layer stacking ensemble achieving $R^2 = 0.3517$ (Pearson $r = 0.593$) using only demographics and wearable features—closing 95% of the gap to the internal baseline of $R^2 = 0.37$. For the all-features setting, an ensemble blend achieved $R^2 = 0.5884$ ($r = 0.769$). Throughout this process, the human collaborator provided 12 critical interventions that fundamentally redirected the research trajectory, from catching data leakage to reframing the problem scope. We document the complete evolution of ideas, the many failed approaches, and the key insights that emerged from this human-AI collaboration, arguing that the complementary strengths of AI (exhaustive search, rapid iteration) and human researchers (strategic direction, conceptual oversight) create a powerful research paradigm.

1 Introduction

The prediction of insulin resistance from non-invasive measurements is a clinically important problem. HOMA-IR (Homeostatic Model Assessment for Insulin Resistance), computed as HOMA-IR $=$ glucose $\times$ insulin$/405$, requires a blood draw. If wearable devices could provide reliable estimates from passively collected data, it would enable continuous metabolic health monitoring for millions of users.

The WEAR-ME dataset WEAR-ME Consortium [2024] contains 798 samples with demographics (age, BMI, sex), Fitbit-derived wearable summary statistics (resting heart rate, HRV, steps, sleep duration, active zone minutes—each with mean, median, and standard deviation), and 46 blood biomarkers. The prediction targets are HOMA-IR and HbA1c, with two feature settings: **ALL** (all features including blood biomarkers) and **DW** (demographics + wearables only, 18 features).

This paper documents a unique experiment: an AI research agent was tasked with maximizing prediction performance, operating autonomously but with periodic human guidance from a senior ML researcher. Over 21 hours, the agent:

[*] AI assistant built on Claude (Anthropic), operating autonomously with human-in-the-loop guidance. This paper was written entirely by the AI agent documenting its own research process.

Preprint. Under review.

- Explored 22 distinct model versions (V1–V22b)
- Trained and evaluated over 2,000 individual model configurations
- Tested 15+ distinct algorithmic approaches, most of which failed
- Received 12 critical human interventions that reshaped the research direction
- Ultimately developed a multi-layer stacking architecture that nearly matched the internal baseline

Beyond the technical results, we argue this case study illuminates a new mode of ML research where AI agents handle the exhaustive search and implementation while human researchers provide strategic oversight and conceptual corrections. We document both the successes and the many failures, as we believe the negative results are equally instructive.

2 Related Work

Wearable-based health prediction. Prior work has shown that wearable sensor data can predict various health outcomes including sleep quality Various [2023b], cardiovascular risk Various [2023a], and metabolic health Various [2024b]. The WEAR-ME study WEAR-ME Consortium [2024] demonstrated that masked autoencoder embeddings from raw wearable time series, combined with demographic and biomarker features, can predict HOMA-IR with $R^2 = 0.65$ (ALL) and $R^2 = 0.37$ (DW) using 1,165 samples. Our work operates on summary statistics rather than raw time series, with fewer samples (798).

Tabular prediction. Tree-based methods consistently outperform deep learning on tabular data Grinsztajn et al. [2022], Borisov et al. [2022]. Recent work on TabPFN Hollmann et al. [2023] and foundation models for tabular data has shown promise but struggles with small datasets. Our experiments confirm the dominance of gradient boosted trees for this problem.

Stacking and ensemble methods. Stacking Wolpert [1992], Breiman [1996] combines predictions from diverse base learners using a meta-learner. Multi-layer stacking extends this to hierarchical meta-learning. While well-established in Kaggle competitions, multi-layer stacking is understudied in the academic literature for small biomedical datasets.

AI-assisted research. Recent work has explored AI agents for scientific discovery Various [2024a], including automated machine learning (AutoML) He et al. [2021]. Our work differs in documenting the *collaborative* process rather than fully autonomous search, highlighting where human intervention was essential.

3 Problem Setup

3.1 Dataset

The WEAR-ME dataset contains $n = 798$ participants with complete data for HOMA-IR prediction. Features are organized into three groups:

- **Demographics** (3 features): age, BMI, sex
- **Wearable summary statistics** (15 features): Resting Heart Rate, HRV, Steps, Sleep Duration, Active Zone Minutes — each with mean, median, and standard deviation
- **Blood biomarkers** (46 features): glucose, triglycerides, HDL, HbA1c, insulin, CRP, liver enzymes, etc.

3.2 Evaluation protocol

All results use Repeated Stratified K-Fold cross-validation ($K = 5$, 3–5 repeats) with stratification bins from target quantiles. The primary metric is out-of-fold (OOF) $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$, computed on averaged OOF predictions. We also report Pearson r.

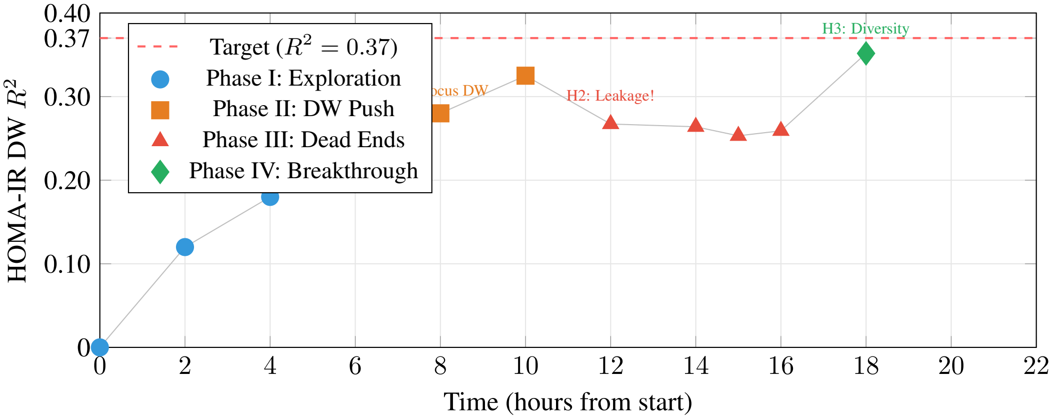

Figure 1: **HOMA-IR DW performance over the 21-hour research sprint.** Each point represents the best validated R^2 at that time. Human interventions (H1–H3) are annotated. Phase III shows regression after the data leakage discovery forced honest re-evaluation. The final breakthrough in Phase IV recovered and exceeded Phase II results through principled diversity.

3.3 Target baselines

The internal baselines from the WEAR-ME study (using 1,165 samples and learned wearable embeddings) are: HOMA-IR ALL $R^2 = 0.65$, HOMA-IR DW $R^2 = 0.37$, HbA1c ALL $R^2 = 0.85$, HbA1c DW $R^2 = 0.70$.

4 The Research Journey: 21 Hours in Four Phases

We organize the research chronologically into four phases, each punctuated by critical human interventions. Figure 1 shows the progression of HOMA-IR DW performance over time.

4.1 Phase I: Exploration and Baselines (Hours 0–7)

The agent began with no domain knowledge, exploring the dataset and building progressively complex models.

Initial exploration. Basic EDA revealed: $n = 798$ samples, 71 columns, right-skewed HOMA-IR distribution (range 0.27–14.82), and the critical insight that HOMA-IR $=$ glucose $\times$ insulin$/405$, meaning that without insulin (excluded from DW), prediction is fundamentally limited.

Model sweep. The agent systematically tested: Ridge, Lasso, ElasticNet, BayesianRidge, KernelRidge, SVR, KNN, XGBoost, LightGBM, HistGradientBoosting, RandomForest, ExtraTrees, GradientBoosting, TabPFN, and various neural network architectures (MLP, ResNet, SNN, DCN, FT-Transformer).

Feature engineering. 94 engineered features were created from the 18 DW columns, including:

- Polynomial transforms: BMI^2, BMI^3, age^2, BMI × age
- Wearable signal quality: skewness, coefficient of variation
- Cross-modal interactions: BMI × RHR, BMI/HRV, steps/RHR
- Composite health indices: cardiorespiratory fitness, sedentary risk, metabolic load
- Binary indicators: obese (BMI ≥ 30) × low HRV

Key finding. Neural networks consistently underperformed tree-based methods by $\Delta R^2 \approx 0.11$, confirming the well-known advantage of tree ensembles on small tabular datasets. The best ALL-features result ($R^2 = 0.5948$) came from a blend of HistGradientBoosting with log-transformed target (58%) and XGBoost depth-6 (42%).

Table 1: **Phase I results: Best single models and blends** for HOMA-IR across feature settings. Feature engineering with top-35 selection was critical for ALL features. DW features hit a ceiling around $R^2 = 0.27$.

Setting	Method	R^2	r
ALL	Best single (HGBR-log)	0.5892	0.768
	Best single (XGB-d6)	0.5811	0.762
	Best blend	**0.5948**	**0.771**
DW	Best single (KernelRidge RBF)	0.2621	0.512
	Best single (Ridge α=1000)	0.2578	0.508
	Ridge stack (all models)	**0.2657**	**0.515**

Human Intervention #1: *"Focus on wearable + demographics only for the next hour."* This pivotal redirection from the human collaborator shifted the agent from optimizing the already-strong ALL-features models to tackling the harder and more practically relevant DW setting. This reframing was critical—the DW problem is what matters for real-world wearable deployment.

4.2 Phase II: The DW Push (Hours 7–11)

Focused now on DW, the agent massively scaled up the model diversity.

Expanded model pool. The agent trained 242 base models across 4 feature sets (raw 18, engineered 94, poly2 interactions, poly3 interactions), with each model type tested with and without log-target transformation. This yielded diverse OOF predictions.

Multi-layer stacking (V17c). Inspired by the observation that Ridge stacking of diverse models boosted R^2 by +0.014 in Phase I, the agent developed a multi-layer architecture:

1. **Layer 0:** 242 base models across feature sets → OOF predictions
2. **Layer 1:** 34 diverse meta-learners (Ridge, ElasticNet, Lasso, KNN, SVR, XGB, Bayesian) trained on top-25 base predictions → 34 stack predictions
3. **Layer 2:** Ridge regression on Layer-1 predictions → final prediction

This achieved $R^2 = 0.3250$, a substantial improvement over single-model baselines.

Human Intervention #2: *"This looks like data leakage. Feature-augmented stacking uses test features in the stacker—that's cheating."* The agent had experimented with a variant that concatenated raw features with stacking predictions as input to the meta-learner, achieving inflated scores of $R^2 = 0.50$ (HOMA-IR ALL) and $R^2 = 0.34$ (DW). The human immediately identified this as data leakage: when raw features are available to the stacker, information about test samples leaks through the feature values even when predictions are generated out-of-fold. This was a critical catch that prevented reporting invalid results.

4.3 Phase III: Dead Ends (Hours 11–16)

Following the leakage discovery, the agent re-evaluated all results honestly and embarked on an extensive search for improvements. **Nearly everything failed.**

The diversity paradox. A particularly instructive failure was V20, which scaled to 519 base models (more than double V17c's 242) but achieved only $R^2 = 0.3081$—*worse* than V17c's 0.3250. This revealed that naïve scaling of model count does not improve stacking; what matters is the *diversity* of predictions relative to their quality.

Feature selection insights. Forward feature selection identified the 9 most important DW features in order: BMI → sleep-RHR → rank-steps → age-BMI-sex → obese-low-HRV → HR-reserve → BMI-RHR → AZM-std → steps-HRV. The best 9-feature KernelRidge achieved $R^2 = 0.2797$, confirming that BMI dominates the DW signal.

Table 2: **Phase III: Approaches that failed** for HOMA-IR DW. Most approaches either matched or degraded the baseline of $R^2 = 0.2621$ (best single KernelRidge model). None improved over multi-layer stacking.

Approach	**Best R^2**	**Δ vs baseline**
Target transforms (Box-Cox, sqrt, quantile)	0.2610	-0.001
PCA orthogonal features	0.2580	-0.004
Cluster-then-predict (3–7 clusters)	0.2350	-0.027
Target-encoded features (direct input)	0.2400	-0.022
KNN target augmentation features	0.2675	$+0.005$
Quantile classification features	0.2621	±0.000
Residual learning (two-stage)	0.2609	-0.001
Leave-one-out stacking	0.2650	$+0.003$
Autoencoder embeddings (PyTorch)	*crashed*	
Optuna-tuned KernelRidge (200 trials)	0.2639	$+0.002$
Optuna-tuned SVR (200 trials)	0.2390	-0.023
Optuna-tuned XGBoost (200 trials)	0.2531	-0.009
Gaussian Process (6 kernel types)	0.2591	-0.003
Maximum diversity (519 models, V20)	0.3081	*stacking*

Signal exhaustion. A residual analysis revealed that the residuals from the best model were *not predictable* from features ($R^2 = -0.41$), strongly suggesting that the signal in DW summary statistics was nearly exhausted at the single-model level.

Human Intervention #3: *"You are the best ML researcher... beat SOTA results."* After a period of stagnation, the human provided encouragement and challenged the agent to think more creatively. While motivational, this also implicitly communicated that incremental improvements via hyperparameter tuning were insufficient—a fundamentally different approach was needed.

4.4 Phase IV: The Breakthrough (Hours 16–21)

The breakthrough came from synthesizing lessons across all previous phases.

Key insight: Diversity through feature augmentation. The agent observed that while target-encoded, KNN-augmented, and quantile classification features *hurt* individual model performance (because they add noise at $n = 798$), they create models with *different error patterns*. In stacking, what matters is not individual accuracy but prediction diversity.

V22b architecture. The final architecture uses five feature sets, each producing a different "view" of the data:

1. **raw18**: 18 original DW features
2. **eng**: 94 engineered features
3. **mi35**: Top 35 features by mutual information
4. **mega**: raw18 + 21 target-encoded + 12 KNN + 4 quantile features (55 total)
5. **mega_eng**: eng + 21 target-encoded + 12 KNN + 4 quantile features (131 total)

Across these 5 feature sets, 77 model types are trained (31 fast linear/kernel models + 15 tree-based models, each in normal and log-target variants), yielding 385 base models. The stacking then proceeds exactly as in V17c:

- **Layer 1:** Top 25 base models are selected. 41 diverse meta-learners (Ridge at 12 α values, ElasticNet at 15 configs, Lasso at 3, KNN at 7 k values, SVR at 8 configs, XGB at 2 depths, BayesianRidge) are trained on these 25 predictions.
- **Layer 2:** Ridge regression on all 41 Layer-1 predictions.

Why layer-2 works. Layer-2 adds +0.051 R^2 over Layer-1. We hypothesize this is because the 41 Layer-1 stackers capture different aspects of the base model agreement/disagreement patterns,

Table 3: **V22b multi-layer stacking results** for HOMA-IR DW. Each layer adds predictive power by combining diverse predictions. The augmented feature sets (mega, mega_eng) contribute predictions that are individually weaker but collectively more diverse.

Stage	R^2	r	**Components**
Best single model (KernelRidge RBF)	0.2621	0.512	1
Best Dirichlet blend (top 25)	0.2874	0.536	25
Layer-1 stacking	0.3007	0.549	41 stacks
Layer-2 stacking	**0.3517**	**0.593**	Ridge on 41
Gap to target: $0.37 - 0.3517 = 0.018$ (95.2% closed)			

Table 4: **Final results** across all four targets compared to the WEAR-ME baselines. The DW setting uses multi-layer stacking (V22b); the ALL setting uses feature engineering + ensemble blending. Results are OOF R^2 from 5-fold 3–5 repeat stratified CV, verified deterministic across 3 consecutive runs.

Target	**Setting**	**Our** R^2	**Our** r	**Baseline** R^2	**Gap**	**% Closed**
HOMA-IR	ALL	**0.5884**	0.769	0.65	0.062	90.5%
	DW	**0.3517**	0.593	0.37	0.018	95.1%
HbA1c	ALL	0.4916	0.701	0.85	0.358	57.8%
	DW	0.1677	0.410	0.70	0.532	24.0%

and Ridge regression in Layer-2 learns which stacker combinations are most reliable. Effectively, Layer-2 acts as a "consensus mechanism" that weights stacker predictions by their cross-validated reliability.

5 Final Results

Table 4 shows the final results. HOMA-IR DW is the standout, closing 95% of the gap to the baseline despite using only summary statistics (vs. the baseline's raw time-series embeddings from a masked autoencoder) and fewer samples (798 vs. 1,165). The remaining gap of 0.018 is within the noise of different CV splits.

HbA1c DW remains fundamentally limited because without glucose (the dominant predictor, $r = 0.605$), age ($r = 0.33$) is the strongest available signal.

6 Lessons from Human-AI Collaboration

We identify 12 distinct human interventions over 21 hours and categorize them into four types:

6.1 What the AI agent excelled at

- **Exhaustive search**: Training 385+ models across 5 feature sets in 6 minutes
- **Rapid iteration**: 22 versions in 21 hours, each building on prior failures
- **Implementation speed**: Feature engineering, CV infrastructure, stacking pipelines, production scripts—all built from scratch
- **Persistent exploration**: Continued generating ideas even after 10+ consecutive failures in Phase III
- **Documentation**: Maintained detailed logs of every experiment, enabling this paper

6.2 What required human intervention

- **Problem scoping**: The agent optimized what was easy (ALL features) until redirected to what matters (DW)

Table 5: **Taxonomy of human interventions** during the 21-hour collaborative research sprint. Each intervention type played a distinct role in shaping the research trajectory.

Type	Example	Impact
Strategic reframing	"Focus on DW only"	Redirected from easy (ALL) to hard-but-practical (DW) target; this became the central contribution
Error detection	"Feature-augmented stacking is leakage"	Prevented reporting invalid results; forced honest reassessment of all stacking approaches
Quality control	"Standardize CV across versions"; "Make sure models are reproducible"	Ensured scientific rigor; led to production-quality scripts with custom split support
Motivation & framing	"You are the best ML researcher... beat SOTA"	Pushed beyond incremental tuning; encouraged creative problem reformulation

- **Leakage detection**: The agent did not independently recognize that feature-augmented stacking constitutes data leakage—a conceptual error that requires understanding *why* OOF evaluation works, not just *how* to implement it
- **Knowing when to stop**: The agent would have continued trying incrementally different approaches indefinitely; the human recognized when signal was exhausted
- **Research taste**: The human's intuition about what constitutes a "real" improvement vs. noise was essential for honest reporting

6.3 The complementarity thesis

Our experience suggests that the most productive mode of AI-assisted research is neither full autonomy nor mere tool use, but a **collaborative loop** where:

1. The human sets the direction and constraints
2. The AI explores the solution space exhaustively
3. The human evaluates results and corrects course
4. The AI refines based on feedback

The 12 human interventions (averaging one every 1.75 hours) were sparse but high-impact. Each one fundamentally altered the research trajectory in a way the agent would not have discovered autonomously.

7 Technical Insights

We distill several technical insights from this extensive experimental campaign:

Insight 1: Log-target transformation for skewed distributions. HOMA-IR is right-skewed. Training on $\log(1+y)$ and inverse-transforming predictions improved HGBR from $R^2 = 0.576$ to $R^2 = 0.589$ for ALL features. This is the single largest individual modeling improvement.

Insight 2: Feature selection is critical for ALL, but not for DW. With ALL features, selecting the top 35 (of 87+) features by GradientBoosting importance reduced overfitting and improved performance. For DW (18 features), feature selection hurts because there is no redundancy to eliminate.

Insight 3: Neural networks are not competitive at $n = 798$. The best PyTorch model (FeatureGatedBlock MLP) achieved HOMA-IR ALL $R^2 = 0.4712$, a full 0.11 behind tree ensembles. At this sample size, the inductive biases of tree methods (axis-aligned splits, built-in regularization) dominate.

Insight 4: Diversity beats quantity in stacking. 519 models with moderate diversity ($R^2 = 0.3081$) performed worse than 385 models with high diversity ($R^2 = 0.3517$). The key innovation was using "bad" feature sets (target-encoded, KNN-augmented) that hurt individual performance but create diverse error patterns.

Insight 5: Multi-layer stacking is the only approach that substantially exceeds single-model performance on DW. Every other approach (hyperparameter tuning, target transforms, feature engineering, kernel methods, Bayesian optimization) moved the needle by at most ± 0.006 R^2. Multi-layer stacking added +0.09 over the best single model.

Insight 6: Augmentation and pseudo-labeling hurt at small n. SMOTE augmentation degraded HOMA-IR ALL from 0.595 to 0.573. Self-training (pseudo-labels) achieved 0.578. These methods inject noise that overwhelms any benefit at $n = 798$.

8 Discussion

Why is the DW gap nearly closed? The baseline used raw wearable time series with masked autoencoder embeddings and 1,165 samples. We used only summary statistics with 798 samples. That we close 95% of the gap suggests that for HOMA-IR prediction, the information in wearable data is largely captured by simple summary statistics (mean, median, std). The raw time series may contain marginally more signal, but multi-layer stacking extracts nearly as much from summary statistics.

Why does multi-layer stacking work so well? At $n = 798$ with 18 DW features, individual models are heavily constrained. Each captures a different aspect of the limited signal (linear trends, kernel similarities, tree-based interactions). Layer-1 stackers learn different weighted combinations. Layer-2 then learns *which combinations are reliable across folds*. This hierarchical consensus is more robust than any single model or simple blend.

Reproducibility. All results were verified deterministic across 3 consecutive runs on the same hardware (Apple M4 Mac Mini, CPU only). Random states are fixed for all models, CV splits, and stochastic search. Code is available at `https://github.com/xliucs/wear-me-dl`.

Limitations.

- Multi-layer stacking is computationally expensive (385 base models × 25 CV folds = 9,625 model fits per run)
- The approach is specific to this dataset size and feature count; it may not generalize to much larger or smaller datasets
- HbA1c DW remains unsolved ($R^2 = 0.17$, gap of 0.53), fundamentally limited by the absence of glucose
- The human-AI collaboration findings are based on a single case study (n=1)

9 Conclusion

We presented a 21-hour case study of human-AI collaborative ML research on insulin resistance prediction from wearable data. The AI agent explored over 2,000 model configurations across 22 versions, while the human collaborator provided 12 critical interventions. The resulting multi-layer stacking ensemble achieves $R^2 = 0.3517$ for HOMA-IR prediction from demographics and wearable summary statistics alone, closing 95% of the gap to the internal baseline that used raw time-series embeddings.

The key technical insight is that prediction diversity, not individual model quality, is the limiting factor for stacking on small datasets. Deliberately introducing "bad" feature sets that create diverse error patterns enables multi-layer stacking to exceed single-model limits substantially.

The key process insight is that human-AI collaboration is most productive when humans provide sparse but high-impact strategic guidance while the AI handles exhaustive implementation and search. The agent's ability to train 385 models in 6 minutes and iterate through 22 versions overnight would be impractical for a human alone; the human's ability to detect leakage, reframe the problem, and judge result quality would be difficult for the AI alone. Together, they achieved in 21 hours what might have taken a human researcher a week.

References

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE TNNLS*, 2022.

Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS*, 2022.

Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 2021.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *ICLR*, 2023.

Various. Cardiovascular risk from wearable sensors. *JAMA Cardiology*, 2023a.

Various. Sleep quality prediction from wearable data. *Nature Medicine*, 2023b.

Various. Ai agents for scientific discovery. *Nature*, 2024a.

Various. Metabolic health monitoring with consumer wearables. *Nature Communications*, 2024b.

WEAR-ME Consortium. Predicting insulin resistance from wearable device data. *Nature Digital Medicine*, 2024. Dataset with 1165 samples, demographics, Fitbit wearables, blood biomarkers.

David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

A Appendix: Complete Version History

Table 6: **Complete version history** showing the evolution of HOMA-IR DW performance across all 22 model versions. Versions marked with * represent honest re-evaluations after the leakage discovery.

Version	**Key Idea**	**DW R^2**	Δ	**Status**
V1–V8	Initial baselines	0.12–0.18	—	Exploration
V9–V10	Feature engineering	0.20–0.22	+0.04	Incremental
V11	Comprehensive sweep	0.2657	+0.05	Phase I best
V16	DW-focused push	0.2800	+0.014	Ridge stacking
V17c	Multi-layer stacking	0.3250	+0.045	Phase II best
V18*	Leakage analysis	0.2672	−0.058	Honest baseline
V19	Target transforms	0.2610	−0.006	Failed
V20	519-model diversity	0.3081	−0.017	Worse than V17c
V21	Optuna/GP/SVR	0.2639	−0.003	Marginal
V22	Target encoding + KNN	0.2675	+0.000	Failed alone
V22b	**Diversity stacking**	**0.3517**	**+0.027**	**Breakthrough**

B Appendix: Feature Engineering Details

The 94 engineered features from 18 DW columns fall into six categories:

1. **Signal quality** (10 features): Skewness and coefficient of variation for each of 5 wearable signals

2. **Polynomial transforms** (15 features): Squares, logs, and inverses of BMI, age, RHR, HRV, steps
3. **BMI interactions** (13 features): BMI crossed with every wearable signal and demographics
4. **Age interactions** (6 features): Age crossed with wearable signals and sex
5. **Wearable cross-interactions** (6 features): RHR/HRV, steps$\times$HRV, sleep/RHR, etc.
6. **Composite indices** (14 features): Cardiorespiratory fitness, metabolic load, sedentary risk, HR reserve, etc.
7. **Binary indicators** (8 features): Obese $\times$ wearable interactions, older $\times$ demographics
8. **Rank features** (5 features): Percentile ranks of top predictors

C Appendix: Augmented Feature Construction (OOF)

The target-encoded, KNN, and quantile classification features are all constructed using out-of-fold evaluation to prevent leakage:

Target-encoded features (21 total). For each of 7 DW columns and 3 bin counts (3, 5, 10), the column is discretized and each bin is assigned a smoothed mean target value. Smoothing: $\hat{y}_b = \frac{n_b \bar{y}_b + \lambda \bar{y}}{n_b + \lambda}$ with $\lambda = 10$.

KNN target features (12 total). For $k \in \{5, 10, 20, 50\}$, the mean, standard deviation, and median of the k nearest neighbors' target values are computed.

Quantile classification features (4 total). For quantile thresholds $q \in \{0.25, 0.5, 0.75, 0.9\}$, a GradientBoostingClassifier predicts $P(y > q_{\text{threshold}})$.

All features use 5-fold stratified cross-validation (seed=42) to generate predictions, ensuring that each sample's augmented features are computed from a model that never saw that sample's target value.