

RESEARCH ARTICLE



WHSDA-AI: Interpretable Walsh-Hadamard Spectral Deviation Analysis for Structured Artificial Intelligence Applications in Banking Sector

Sergo A. Episkoposian¹ and Suren H. Parsyan^{2,*}

¹*Faculty of Applied Mathematics and Physics, National Polytechnic University of Armenia, Armenia*

²*Service-Based Business Management, Armenian State University of Economics, Armenia*

Abstract: Artificial intelligence (AI) in banking increasingly requires models and representations that are not only predictive but also interpretable, auditable, and stable under model updates. Fraud screening, anti-money-laundering triage, and customer-risk monitoring are high-stakes structured-data tasks in which explanations must be reproducible and connected to documented feature processing. This paper proposes WHSDA-AI, a Walsh-Hadamard Spectral Deviation Analysis framework for structured and profile-based AI in the banking sector and related tabular risk-analysis settings. The proposed framework maps normalized feature profiles into a fixed orthogonal signed-contrast space. Each retained coefficient represents a deterministic comparison between two documented groups of input features. The novelty of the approach is not the classical Walsh-Hadamard transform itself but its integration into a governance-oriented AI layer: semantic feature-map versioning, a fixed coefficient dictionary, reference-normalized deviation analysis, leakage-free coefficient selection, bootstrap stability assessment, and compact explanation-card generation. The full Walsh-Hadamard transform is an isometry and cannot improve Euclidean rankings by itself. Practical value arises through coefficient selection, reference deviations, template comparison, risk weighting, and integration with supervised models. Experiments on banking-style diagnostic data, credit-style fraud data, and public structured-data benchmarks compare WHSDA-AI with raw-feature models, sparse logistic regression, principal component analysis (PCA), Haar contrasts, random signed projections, random forests, gradient boosting, XGBoost, and LightGBM. Results are reported with repeated-split uncertainty and conservative robustness interpretation. The framework is positioned as an ante-hoc representation, governance, and feature-engineering layer rather than as a universal accuracy booster.

Keywords: machine learning, fraud detection, banking risk, XGBoost, LightGBM

1. Introduction

Modern artificial intelligence (AI) is increasingly deployed in settings where predictive performance alone is not sufficient. In banking, healthcare, industrial monitoring, cybersecurity, and public-sector decision support, a model must often justify why an object is ranked as risky, anomalous, or clinically relevant. In banking applications, this requirement is especially acute because fraud screening, anti-money-laundering (AML) triage, customer-risk profiling, and transaction-risk monitoring are examined by compliance officers, auditors, and model-risk teams. Explanation must therefore be stable enough to survive model updates and transparent enough to be connected with documented feature processing [1–3]. The increasing integration of AI-powered virtual assistants into banking service frameworks further illustrates

the growing operational penetration of AI across customer-facing financial services [4].

The need for interpretability has led to a wide range of post hoc explanation methods, including SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), counterfactual explanations, and model-level interpretability handbooks [5–8]. These tools are valuable, but they usually explain a trained model after the representation has already been formed. In parallel, recent ante-hoc interpretability research emphasizes models and representations whose explanatory structure is visible before deployment [9–11].

At the same time, structured tabular data remain central in applied machine learning. TabNet, deep tabular architectures, retrieval-augmented TabR, TabReD, TabArena, TabFSBench, and TabPFN have substantially expanded the modern tabular-learning landscape [12–23]. However, several benchmark studies continue to show that no single model family dominates all tabular tasks and that evaluation protocols must be carefully controlled [15–20]. Tree-based models and boosted trees

*Corresponding author: Suren H. Parsyan, Service-Based Business Management, Armenian State University of Economics, Armenia. Email: parsyan.suren@asue.am

remain highly competitive in practical tabular learning, including XGBoost and related gradient-boosting approaches [15–17, 24].

This paper addresses a complementary question: can one introduce a simple, fixed, auditable representation layer that produces interpretable feature-group contrasts before or alongside a supervised model? We propose WHSDA-AI, a Walsh-Hadamard Spectral Deviation Analysis framework for structured AI profiles. The method is not a new black-box classifier. It is a transparent feature-engineering, diagnostic, and explanation layer that can be inserted into existing machine-learning pipelines and compared with raw features, principal component analysis (PCA), Haar contrasts, random projections, tree-based models, tabular deep models, and post hoc explanations [5–11, 22–30].

The central mathematical fact is deliberately not hidden. The normalized Walsh-Hadamard transform (WHT) is orthogonal. Therefore, if all coefficients are retained and the same Euclidean metric is used, the ranking in WHT space is identical to the ranking in the original normalized space. The method can add value only when selected coefficients, weighted contrasts, reference profiles, suspicious or target templates, or downstream nonlinear models are used. This conservative positioning prevents exaggerated claims and gives a strict implementation sanity check [26–28, 31].

Relative to similar work, WHSDA-AI differs in its object of explanation. SHAP, LIME, and counterfactual tools explain a fitted model; PCA and random projections provide alternative coordinates; gradient-boosted trees provide strong prediction; Explainable Boosting Machine (EBM)-style models provide fitted glass-box feature effects. WHSDA-AI instead provides a fixed, versioned signed-contrast dictionary and a reference-deviation explanation layer whose coefficient meanings are defined before model fitting. This distinction is central for banking-sector audit trails because the same feature-map version and coefficient dictionary can regenerate the explanation independently of a particular classifier.

This paper makes six contributions. First, it formulates WHT signed contrasts as a general AI feature layer for structured banking and tabular risk data rather than as a purely descriptive visualization. Second, it introduces reference-normalized spectral deviation, allowing each object to be compared with normal, target, or suspicious prototypes. Third, it provides a leakage-free evaluation protocol with coefficient selection, bootstrap stability, isometry checks, raw-versus-raw-plus-WHT ablation, and repeated-split uncertainty. Fourth, it reports comparative results against sparse logistic regression, PCA, Haar contrasts, random signed projections, random forests, gradient boosting, XGBoost, and LightGBM. Fifth, it clarifies the relation of WHSDA-AI to EBM-style interpretable additive models and TabPFN-style tabular foundation models, while explaining why modern boosted-tree systems are the closest production-grade comparators for the reported tabular risk-ranking objective. Sixth, it integrates the method with AI governance expectations, including NIST AI RMF, ISO/IEC 42001:2023, and the EU AI Act [1–3].

2. General AI Formulation of the Contrast-Layer Approach

Let an object in an AI task be represented by a structured feature profile. The object may be a bank account, a customer, a merchant, a patient, a machine, a network host, a document-level feature profile, a graph node, or an image after feature extraction. The only requirement is that features can be organized into

meaningful blocks or at least documented as a stable ordered vector. In the banking case, such blocks include activity, volume, balance dynamics, counterparty structure, velocity, risk ratios, graph information, and device evidence.

$$x = (x_1, x_2, \dots, x_p) \in \mathbb{R}^p. \quad (1)$$

After train-only normalization and optional padding, the profile is represented as a vector z in a power-of-two dimension n . Padding coordinates are inactive, are appended only after normalization, and are recorded in the coefficient dictionary.

$$z = \text{Normalize}_{\text{train}}(x) \in \mathbb{R}^n, \quad n = 2^m, \quad n \geq p. \quad (2)$$

Let H_n be the normalized Walsh-Hadamard matrix of order n . Its rows are denoted by $h_0, h_1, \dots, h_{n-1}$. The normalization is chosen so that H_n is orthogonal.

$$H_n H_n^T = I_n, \quad H_n^T H_n = I_n. \quad (3)$$

The Walsh-Hadamard coefficient vector is obtained by applying this fixed signed dictionary to the normalized profile.

$$c(z) = H_n z = (c_0(z), c_1(z), \dots, c_{n-1}(z)). \quad (4)$$

The k -th coefficient is a signed contrast between two groups of coordinates determined by the signs in the k -th row of H_n . If s_{kj} is the sign of the j -th coordinate in row k , then the positive and negative feature groups are fixed before training.

$$P_k = \{j : s_{kj} = +1\}, \quad N_k = \{j : s_{kj} = -1\}. \quad (5)$$

$$c_k(z) = \langle h_k, z \rangle = \frac{1}{\sqrt{n}} \left(\sum_{j \in P_k} z_j - \sum_{j \in N_k} z_j \right). \quad (6)$$

Because the transform is orthogonal, the full WHT preserves inner products and Euclidean distances. This property is central: a full-WHT representation cannot by itself improve Euclidean distance ranking. Any empirical gain must come from coefficient selection, reference deviations, weighting, semantic grouping, or nonlinear downstream interactions.

$$\|H_n z - H_n u\|_2 = \|z - u\|_2, \quad \langle H_n z, H_n u \rangle = \langle z, u \rangle. \quad (7)$$

A coefficient-selection operator extracts a stable subset S_K of K coefficients. The selected vector may be appended to raw features, used to compare the object with reference prototypes, or translated into an explanation card. Because H_n is fixed, the meaning of each retained coefficient can be stored in a coefficient dictionary before any classifier is fitted.

$$S_K \subset \{0, 1, \dots, n-1\}, \quad \phi_S(z) = (c_k(z))_{k \in S_K}. \quad (8)$$

The complete domain-agnostic WHSDA-AI workflow is illustrated in Figure 1. Representative application domains, structured object profiles, possible semantic feature blocks, and the corresponding interpretation of Walsh-Hadamard contrasts are summarized in Table 1.

Figure 1
Domain-agnostic WHSDA-AI pipeline for auditable structured AI profiles
WHSDA-AI pipeline for auditable structured AI profiles

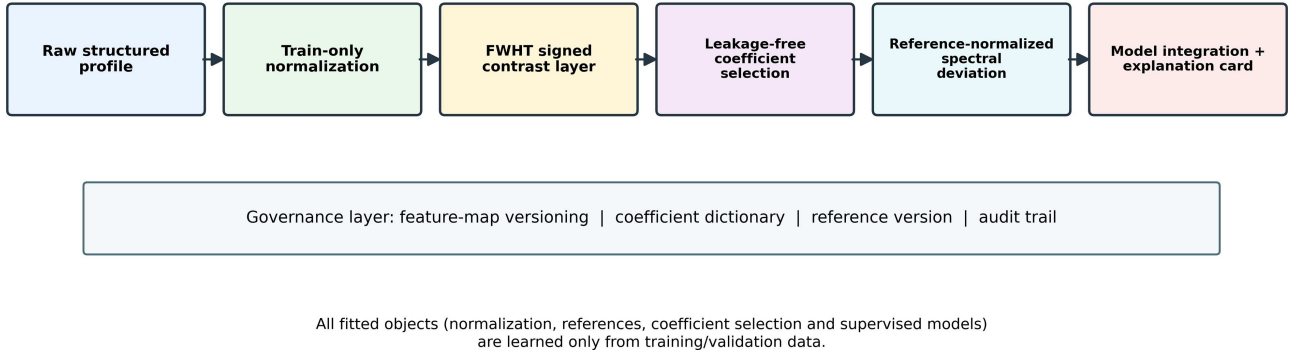


Table 1
General applicability of Walsh-Hadamard contrast layers beyond banking

Domain	Object profile	Possible feature blocks	Interpretation of contrasts
Banking and AML	Account/customer profile	Volume, velocity, balance, counterparty, risk ratios	Transaction-behavior imbalance
Medical diagnosis	Patient profile	Biomarkers, symptoms, imaging-derived features	Physiological deviation from a reference
Industrial monitoring	Machine or sensor profile	Vibration, temperature, pressure, load	Abnormal operating regime
Cybersecurity	Host or network-session profile	Packets, ports, failed logins, duration	Behavioral anomaly signature
Text analytics	Document-level feature profile	Topics, sentiment, lexical/syntactic features	Semantic or stylistic imbalance
Computer vision	Extracted feature vector	Texture, edge, region, embedding blocks	Visual-structure contrast

3. Reference-Normalized Spectral Deviation

The basic WHT coefficient vector describes an object in a fixed contrast basis. A more useful explanatory layer compares the object with a documented reference. Let r be a reference profile, such as a normal prototype, class centroid, suspicious template, clinical baseline, or machine-normal operating state. The reference must be constructed from training data only or from externally approved expert knowledge. It must never use test labels or test distributional statistics.

$$\Delta_k(z; r) = c_k(z) - c_k(r) = \langle h_k, z - r \rangle. \quad (9)$$

$$\Delta_S(z; r) = (\Delta_k(z; r))_{k \in S_K}. \quad (10)$$

In the banking case, two references are especially useful: a normal reference r_0 and a suspicious or target reference r_1 . The first vector explains how the object deviates from normal behavior. The second explains how the object differs from or approaches a target suspicious pattern.

$$\begin{aligned} r_0 &= \text{median}_{\text{train}}\{z_i : y_i = 0\}, \\ r_1 &= \text{median}_{\text{train}}\{z_i : y_i = 1\}. \end{aligned} \quad (11)$$

A simple template score can be defined as a weighted selected deviation score. The weights may come from expert risk priorities,

validation performance, or bootstrap-stable coefficient selection. The manuscript uses this scoring idea only as one component; the selected coefficients can also be appended to supervised models.

$$\text{Score}(z; r_0) = \sum_{k \in S_K} w_k |\Delta_k(z; r_0)|, w_k \geq 0. \quad (12)$$

This property is important in regulated environments because references, feature maps, and coefficient dictionaries must be reproducible across audits [1–3]. If a reference changes after deployment, the change must be versioned together with the feature map and the model card.

4. Relation to Interpretable AI, Tabular Learning, and Modern Benchmarks

WHSDA-AI is not a replacement for SHAP, LIME, counterfactual explanations, concept bottleneck models, boosted trees, or learned tabular representations. It occupies a different position: it produces a fixed and auditable representation before or alongside model training. SHAP and LIME explain a model prediction; WHSDA-AI explains the structural contrast pattern of an object [5–8]. Concept bottleneck models learn or use human-level concepts as intermediate representations [9]. WHSDA-AI does not require concept labels, but it does require a stable semantic feature map.

The distinction between post hoc, ante-hoc, and representation-level interpretability has become more important in recent surveys. Ante-hoc methods attempt to build interpretability into the model or its representation before deployment, while tabular Explainable Artificial Intelligence (XAI) studies highlight the special importance of transparent explanations for structured data in finance, healthcare, and enterprise applications [10, 11]. WHSDA-AI belongs to this representation-level family, but it remains deliberately simple: it uses a fixed orthogonal dictionary rather than a learned concept space.

Modern tabular learning now includes classical gradient-boosted decision trees (GBDT) methods, XGBoost, LightGBM, CatBoost, deep tabular architectures, retrieval-augmented tabular models, and tabular foundation models [12–21, 24]. These methods compete primarily on predictive performance. WHSDA-AI is different: it is a plug-in representation and explanation layer. It can be appended to raw features and used with these predictors, but its main output is an interpretable deviation pattern rather than a stand-alone black-box prediction.

Orthogonal representations such as PCA, Haar transforms, random projections, and Hadamard-based features are also natural competitors [26–30]. PCA is data-adaptive and can be statistically efficient, but its components can drift when the training distribution changes. Haar contrasts are local and hierarchical, which is useful when variables have a natural order. Recent hybrid wavelet-decomposition architectures further demonstrate that structured multiscale transform representations continue to be investigated in contemporary artificial intelligence systems [32]. Random signed projections are cheap, but they lack a stable semantic dictionary. WHSDA-AI uses a fixed Walsh-Hadamard signed dictionary because its rows provide deterministic binary partitions of feature coordinates into positive and negative groups.

For the same-objective empirical comparison, the manuscript separates predictive competitiveness from explanation-dictionary stability. XGBoost and LightGBM are included because they are widely used high-performance tabular models for ranking, fraud screening, and structured risk analysis [24, 33]. Recent comparative studies of gradient-boosted methods further confirm that

XGBoost, LightGBM, and CatBoost remain among the strongest practical competitors for structured tabular classification tasks [34]. EBM is acknowledged as an interpretable additive-model comparator, while TabPFN is acknowledged as a modern tabular foundation-model comparator [21, 35]. Recent EBM-based work in insurance risk modeling also confirms that glass-box additive boosting models remain important comparators when predictive accuracy must be balanced with transparent feature-effect interpretation [36]. WHSDA-AI is not positioned as a replacement for these methods. Its distinct contribution is the fixed signed coefficient dictionary and the reference-deviation explanation layer, which can be used before or alongside such predictors.

The principal methodological differences between WHSDA-AI and these post hoc, ante-hoc, representation-learning, and tabular-prediction approaches are summarized in Table 2.

5. Banking Case Study: Transaction-Risk Monitoring

Although the proposed method is domain-agnostic, banking transaction-risk monitoring is a natural case study because fraud and AML alerts require both ranking and explanation. The monitored object may be an account, customer, merchant, wallet, corporate entity, or counterparty. Raw transactions are aggregated over a fixed observation window into activity, volume, balance, counterparty, velocity, graph, device, and risk-ratio features. Classical and modern fraud-detection literature highlights the importance of class imbalance, anomaly detection, graph structure, and explainability in this setting [24, 25, 38–46].

The banking use case is treated as a case study rather than as the exclusive scope of the method. The same contrast-layer logic applies whenever an object is represented by structured features and where explanation benefits from comparing groups of variables rather than isolated coordinates. Public transaction-risk and credit-card fraud datasets can be used for the next external empirical stage, while institutional validation must follow data-protection and governance requirements [1–3, 43–46].

Table 2
Position of WHSDA-AI among interpretable AI and representation-learning methods

Approach	Main idea	Difference from WHSDA-AI
SHAP	Post hoc additive feature attribution	Explains a trained model rather than a fixed representation [5]
LIME	Local surrogate model around one prediction	Local and model-dependent [6]
Counterfactual explanations	Minimal changes needed to alter a decision	Action-oriented but not a fixed contrast space [7]
Concept bottleneck models	Prediction through human concepts	Requires concept labels or concept supervision [9]
PCA	Data-adaptive orthogonal components	May drift and may lack semantic stability [30]
Haar contrasts	Local hierarchical differences	Requires meaningful local ordering of variables [29]
Random signed projections	Cheap compressed features	No stable semantic interpretation
XGBoost/LightGBM/CatBoost	High-capacity gradient-boosted tree prediction	Strong modern tabular predictors, but no fixed signed coefficient dictionary or reference-deviation explanation layer [24, 35, 37]
TabNet/TabR/TabPFN	Learned or foundation-style tabular prediction	Powerful modern tabular predictors, but not designed as fixed auditable signed-contrast dictionaries [12–21]
WHSDA-AI	Fixed signed group contrasts and reference deviations	Model-independent, auditable, cheap, and contrast-based

The banking-oriented feature blocks used to organize the transaction-risk case study, together with representative variables and their financial interpretations, are summarized in Table 3.

6. Methodology

6.1. Leakage-free preprocessing

The train-validation-test split is performed before normalization, reference construction, coefficient selection, PCA fitting, or supervised model training. Normalization parameters are estimated only on the training set and applied unchanged to validation and test profiles. Test labels are used only once for final evaluation. This protocol follows the reproducibility logic of modern benchmark suites and avoids leakage problems emphasized in industrial tabular benchmarks [18–23].

$$\mu_j = \text{mean}_{\text{train}}(x_j), \sigma_j = \text{sd}_{\text{train}}(x_j), z_j = \frac{x_j - \mu_j}{\sigma_j + \varepsilon}. \quad (13)$$

If the number of features is not a power of two, the profile is padded with inactive neutral coordinates after normalization. Padding coordinates are recorded in the coefficient dictionary and are never interpreted as business signals.

6.2. Semantic feature map and coefficient dictionary

Because coefficient meaning depends on feature order, feature ordering is treated as a governance object. Variables are grouped into stable semantic blocks, ordered within each block

by a documented rule, versioned, and stored with the model card. When variables are added or removed, the feature map receives a new version number, and the coefficient dictionary is regenerated.

For every retained coefficient, the dictionary stores the coefficient index, positive-side features, negative-side features, block membership, reference association, selection frequency, validation score, and a human-readable interpretation. This is the main difference between WHSDA-AI and a purely numerical WHT feature transform. The resulting governance protocol for semantic feature ordering, padding, feature-map versioning, and coefficient-dictionary maintenance is summarized in Table 4.

6.3. Coefficient selection

The retained coefficient set is selected without test leakage. Candidate values of K are evaluated on validation data only. The final K and the coefficient dictionary are frozen before test evaluation. The selection rule should prefer the smallest statistically comparable K because interpretability degrades when many coefficients are retained. Average precision (AP), Area Under the Precision-Recall Curve (PR-AUC), and Area Under the Receiver Operating Characteristic Curve (ROC-AUC) are therefore reported jointly because their behavior under severe class imbalance depends on prevalence and on the structure of classification errors; Area Under the Precision-Recall Curve (AUPRC) should not be regarded as universally superior to Area Under the Receiver Operating Characteristic Curve (AUROC) solely because the positive class is rare [25, 38].

$$S_K = \text{TopK}_k(Q_k), \quad K \in \{4, 8, 16, 32\} \quad (14)$$

Table 3
Banking feature blocks used in the transaction-risk case study

Feature block	Typical variables	Financial interpretation
Activity	Transaction count; active days; off-hour activity; burst frequency	Abnormal intensity or sudden behavior change
Volume	Total amount; outgoing amount; maximum amount; average amount	High exposure or unusual payment scale
Balance dynamics	Initial/final balance; balance drop; cash-out ratio	Possible account draining or short-term laundering behavior
Counterparty structure	Unique recipients; repeated recipients; counterparty concentration	Mule-account, layering or collusive patterns
Velocity	Transactions per hour; short-window repetition	Fast execution of suspicious transfer patterns
Risk ratios	Outgoing/incoming ratio; failed transactions; high-risk share	Structural imbalance relative to normal behavior
Graph and device	In-degree; out-degree; neighborhood risk; device diversity	Network-based and channel-based suspicion signals

Table 4
Governance protocol for stable WHSDA-AI interpretation

Step	Action	Purpose
Define feature blocks	Group variables into semantic blocks	Translate contrasts into domain language
Order blocks	Follow a stable task taxonomy	Keep coefficient meaning stable across experiments
Order features within blocks	Use time scale, amount scale, expert rule, or risk priority	Avoid arbitrary coefficient interpretations
Pad after block completion	Mark inactive coordinates	Prevent artificial contrast explanations
Version feature map	Store version, normalization policy and dictionary	Support auditability and reproducibility

$$K^* = \arg \max_K AP_{\text{validation}}(S_K), \quad (15)$$

ties are resolved by the smaller K .

Coefficient stability is computed by a genuine bootstrap-overlap procedure rather than by a simulated stability score. If S_K is the selected set on the original training split and $S_K^{(b)}$ is the set selected on the b -th bootstrap sample, the reported stability is

$$\text{Stability}(S_K) = \frac{1}{B} \sum_{b=1}^B \frac{|S_K \cap S_K^{(b)}|}{K}. \quad (16)$$

The coefficient-selection mechanism should also be distinguished from Group Lasso and Sparse Group Lasso. Those methods estimate predictive models with group-based sparsity penalties and are useful when the goal is to select variables or groups of variables [47, 48]. Recent grouped-feature-selection literature confirms the continuing relevance of exploiting feature groups in high-dimensional learning [37]. WHSDA-AI addresses a different representational problem: it constructs fixed signed contrasts between documented positive-side and negative-side feature groups and then selects stable contrast coefficients for explanation and downstream modeling. The leakage-free coefficient-selection strategies considered in WHSDA-AI and their intended roles are summarized in Table 5.

6.4. Reproducible algorithm

The complete reproducible WHSDA-AI procedure, including preprocessing, coefficient construction, validation-only selection,

benchmarking, and explanation-card generation, is summarized step by step in Table 6.

Figure 2 illustrates a representative validation-only selection of the retained coefficient number, emphasizing that it is chosen and frozen before the test partition is evaluated.

7. Experimental Design

The evaluation is designed to validate the mechanism conservatively and to address the main threats to empirical interpretation. The first layer is a controlled banking diagnostic benchmark that tests whether WHSDA-AI behaves as expected when suspicious profiles are generated by structured imbalances among feature blocks. The second risk-style layer is a credit-style fraud benchmark with rare positives and grouped deviations. The third layer uses standard structured-data benchmarks from scikit-learn to demonstrate that the method is not limited to banking terminology. These benchmarks are not presented as a universal leaderboard claim; they are used to validate the mechanism, compare representation choices under controlled splits, and test whether the proposed contrast layer remains interpretable when predictive gains are small [22, 23, 49, 50]. The individual empirical validation layers, their datasets, methodological roles, and current study status are summarized in Table 7.

To provide a contemporary comparative study rather than a comparison with simple baselines only, the experimental design includes logistic regression, sparse logistic regression, random forest, gradient boosting, PCA plus gradient boosting, Haar plus gradient boosting, random signed projections plus gradient

Table 5
Leakage-free coefficient-selection rules

Rule	Procedure	Use
Reference-energy selection	Retain the smallest K explaining a fixed share of reference-deviation energy	Weakly supervised ranking and explanation
Validation AP selection	Choose K maximizing validation AP; break ties by smaller K	Supervised raw-plus-WHT models
Bootstrap stability	Retain coefficients selected frequently across resamples	Stable explanations and governance
Interpretability penalty	Prefer shorter and clearer explanations when metrics are statistically similar	High-stakes human review

Table 6
WHSDA-AI algorithmic protocol

Step	Operation	Leakage control/output
1.	Split data into train, validation, and test	No preprocessing before split
2.	Fit normalization on train and transform all splits	Store normalization parameters
3.	Pad to the nearest power of two after normalization	Record inactive coordinates
4.	Build WHT coefficients with FWHT	No labels required for transform
5.	Build references from train data or external expert templates	Store reference version
6.	Rank/select coefficients on train/validation only	Freeze K and coefficient dictionary
7.	Train raw, WHT-only, raw + WHT, PCA, Haar, random sign, and boosted-tree models	Use identical splits and metrics
8.	Run isometry sanity check with full-WHT Euclidean distances	Expected max difference close to machine precision
9.	Evaluate once on test for each split	Report AP, PR-AUC, ROC-AUC, P@20, and uncertainty
10.	Generate explanation card	Include coefficient meanings, references, and model consistency

Figure 2
Validation-only coefficient selection: K is chosen before test evaluation

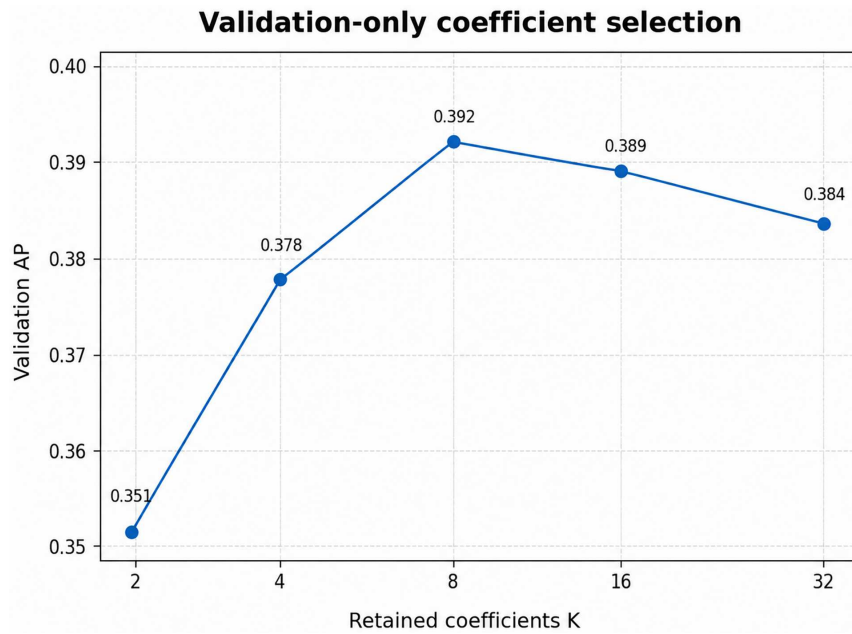


Table 7
Empirical validation layers in the study

Layer	Datasets/tasks	Role in the paper	Status
Diagnostic banking	1200 simulated account profiles; 64 features; 4% positives	Controlled structural-imbalance mechanism test	Reported
Credit-style fraud	1200 profiles; 48 features; 4% rare positives	Fraud-like imbalanced benchmark with grouped deviations	Reported
Small public structured data	Breast Cancer Wisconsin, Wine class-0-vs-rest, Digits high-vs-low	Minimal non-banking reproducibility check	Reported
OpenML/OpenML-CC18	Curated public tabular tasks	Future generalization stress test beyond the reported mechanism validation	External robustness extension
TabReD/TabArena/TabFSBench	Industrial temporal, living and feature-shift benchmarks	Future temporal-drift and feature-shift stress test	External robustness extension
External fraud-data extension	Public/licensed transaction-risk and fraud datasets	Future public fraud/AML validation track	External fraud-data extension

boosting, selected WHT plus gradient boosting, raw plus selected WHT, XGBoost, and LightGBM. The earlier phrase “group-sparse logistic” is not used because the implemented comparator is an L1-regularized sparse logistic regression rather than a true Group Lasso model. True Group Lasso and Sparse Group Lasso remain compatible supplementary baselines [47, 48].

The comparator set is deliberately organized into four families. The first family tests whether WHSDA-AI adds anything beyond raw variables. The second tests whether the gain is merely an orthogonal-transformation effect by comparing against PCA, Haar contrasts, and random signed projections. The third tests modern tabular predictive strength through XGBoost and LightGBM. This choice is supported by recent benchmark work showing that gradient-boosted tree systems, including XGBoost, LightGBM, and CatBoost, remain among the strongest practical methods for structured tabular classification [24, 33, 34]. The fourth evaluates the explanation layer through compactness, bootstrap stability, fidelity, and dictionary coverage. This design

directly addresses the technical question of whether the proposed Walsh-Hadamard contrast layer is different from existing representations and whether its reported benefit survives comparison with stronger contemporary methods [51].

For each benchmark, the primary metrics are AP, PR-AUC, ROC-AUC, and Precision@20. Repeated stratified splits are used to reduce dependence on a single partition. Interpretability is measured through compactness K, bootstrap stability, fidelity, and dictionary coverage. Fidelity is defined as the top-decile overlap between the WHT-only explanation score and the raw-plus-WHT hybrid model.

$$\text{Fidelity} = \frac{|\text{Top}_{10\%}(\text{Score}_{\text{WHT}}) \cap \text{Top}_{10\%}(\text{Score}_{\text{hybrid}})|}{|\text{Top}_{10\%}|}. \quad (17)$$

The complete comparator families and the specific methodological purpose of each baseline are summarized in Table 8.

Table 8
Comparator families used in the experiments

Comparator	Purpose	Relation to WHSDA-AI
Logistic raw	Linear baseline	Tests whether WHT adds anything beyond raw linear information
Sparse logistic regression	Sparse interpretable selection	Corrected label for the L1-regularized comparator
Random forest/gradient boosting	Tree-based tabular predictors	Strong practical nonlinear baselines
XGBoost/LightGBM	Modern boosted-tree comparators	Responds to the reviewer request for methods beyond simple baselines; recent benchmarking confirms the continued strength of gradient-boosted trees on structured tabular classification [24, 33, 34]
PCA + gradient boosting	Adaptive orthogonal representation	Tests against data-adaptive components [30]
Haar + gradient boosting	Alternative fixed orthogonal representation	Tests hierarchical contrasts [29]
Random signs + gradient boosting	Cheap random projection control	Tests whether semantic sign patterns matter
Selected WHT + gradient boosting	Interpretable contrast-only model	Measures retained contrast information
Raw + selected WHT + boosted model	Hybrid WHSDA predictor	Tests whether selected contrasts add useful variables
EBM/TabPFN	Interpretable additive model and tabular foundation-model references	Interpretable additive model and tabular foundation-model references Acknowledged contemporary comparators; recent EBM-based risk-modeling work shows the relevance of glass-box additive boosting, while WHSDA-AI differs by providing a fixed signed-contrast dictionary rather than a fitted additive surface or foundation-model prediction [21, 35, 36].

8. Results

The results are reported in a conservative way. WHSDA-AI is evaluated as a representation and explanation layer, not as a claim that WHTs universally improve all classifiers. The important tests are therefore whether selected WHT contrasts outperform random sign dictionaries and Haar/PCA controls in structured risk settings, whether raw-plus-WHT features can complement modern boosted-tree predictors, and whether the explanation layer remains compact and stable.

8.1. Banking diagnostic benchmark

The banking diagnostic benchmark is not external banking validation. It is a controlled test of the pipeline. In this setting, selected WHT contrasts are expected to help because the positive class is generated through group-level imbalance rather than arbitrary coordinate noise. In the repeated-split version, sparse logistic regression achieves $AP = 0.654 \pm 0.143$, raw XGBoost achieves $AP = 0.301 \pm 0.131$, and raw LightGBM achieves $AP = 0.306 \pm 0.074$. Selected WHT plus gradient boosting reaches $AP = 0.619 \pm 0.117$, while raw plus selected WHT plus XGBoost reaches $AP = 0.643 \pm 0.071$. These results show that WHSDA-AI is competitive in the structured-imbalance regime and can complement modern boosted-tree predictors. The repeated-split average-precision comparison is visualized in Figure 3, while the complete banking diagnostic results for AP, PR-AUC, ROC-AUC, Precision@20, and the retained coefficient count are reported in Table 9.

The representation ablation shown in Figure 4 provides a direct comparison of selected WHT coefficients with PCA, Haar, and random signed projections under the same gradient-boosting setting.

8.2. General structured AI benchmarks

The credit-style fraud benchmark provides a second risk-style test. Logistic and sparse logistic models remain strong because the simulated fraud signal has partially linear group-level structure. The selected WHT representation remains informative and is stronger than PCA, Haar, and random signed projections. Raw plus selected WHT plus XGBoost reaches $AP = 0.532 \pm 0.148$, improving over raw XGBoost $AP = 0.476 \pm 0.185$ in this benchmark. The repeated-split average-precision comparison for this credit-style fraud experiment is shown in Figure 5.

The corresponding complete metric results across all evaluated models are reported in Table 10.

The general public benchmarks show a more nuanced pattern. On saturated datasets such as Breast Cancer and Wine, strong raw-feature models already achieve very high AP, so the selected WHT layer has little room to improve accuracy. On Digits high-vs-low, selected WHT contrasts are competitive but do not replace strong raw-feature random forests. These results are important: WHSDA-AI is not universally accuracy-improving. Its empirical benefit depends on whether class separation is meaningfully expressed through stable group-level contrasts. This heterogeneous pattern across the Breast Cancer, Wine, and Digits benchmarks is summarized in Table 11.

Figure 3
Banking diagnostic benchmark: repeated-split AP for raw and selected-WHT variants

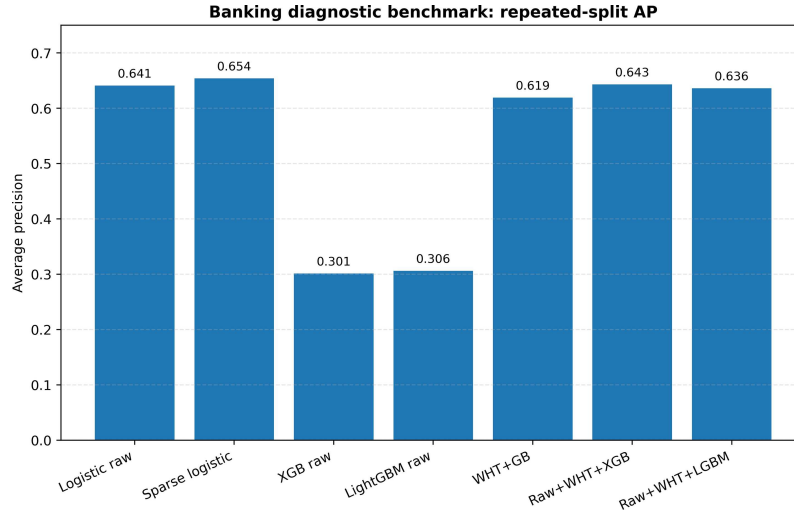


Table 9
Banking diagnostic benchmark results, mean $\pm$ standard deviation over five splits

Model	AP	PR-AUC	ROC-AUC	P@20	K
Logistic raw	0.641 $\pm$ 0.139	0.632 $\pm$ 0.142	0.938 $\pm$ 0.054	0.390 $\pm$ 0.074	4.0
Sparse logistic regression	0.654 $\pm$ 0.143	0.644 $\pm$ 0.145	0.946 $\pm$ 0.049	0.380 $\pm$ 0.057	4.0
Random forest raw	0.248 $\pm$ 0.085	0.225 $\pm$ 0.088	0.814 $\pm$ 0.046	0.210 $\pm$ 0.082	4.0
Gradient boosting raw	0.135 $\pm$ 0.048	0.131 $\pm$ 0.065	0.686 $\pm$ 0.071	0.120 $\pm$ 0.045	4.0
XGBoost raw	0.301 $\pm$ 0.131	0.276 $\pm$ 0.132	0.815 $\pm$ 0.120	0.260 $\pm$ 0.096	4.0
LightGBM raw	0.306 $\pm$ 0.074	0.282 $\pm$ 0.079	0.805 $\pm$ 0.082	0.220 $\pm$ 0.027	4.0
PCA + Gradient boosting	0.185 $\pm$ 0.096	0.174 $\pm$ 0.095	0.799 $\pm$ 0.119	0.180 $\pm$ 0.091	4.0
Haar + Gradient boosting	0.184 $\pm$ 0.105	0.165 $\pm$ 0.092	0.784 $\pm$ 0.125	0.140 $\pm$ 0.119	4.0
Random signs + Gradient boosting	0.085 $\pm$ 0.020	0.077 $\pm$ 0.018	0.677 $\pm$ 0.075	0.080 $\pm$ 0.057	4.0
Selected WHT + Gradient boosting	0.619 $\pm$ 0.117	0.611 $\pm$ 0.131	0.954 $\pm$ 0.044	0.420 $\pm$ 0.084	4.0
Raw + selected WHT + XGBoost	0.643 $\pm$ 0.071	0.620 $\pm$ 0.057	0.952 $\pm$ 0.050	0.430 $\pm$ 0.097	4.0
Raw + selected WHT + LightGBM	0.636 $\pm$ 0.138	0.623 $\pm$ 0.142	0.937 $\pm$ 0.073	0.430 $\pm$ 0.091	4.0

Figure 4
Representation ablation: selected WHT against PCA, Haar, and random signed projections

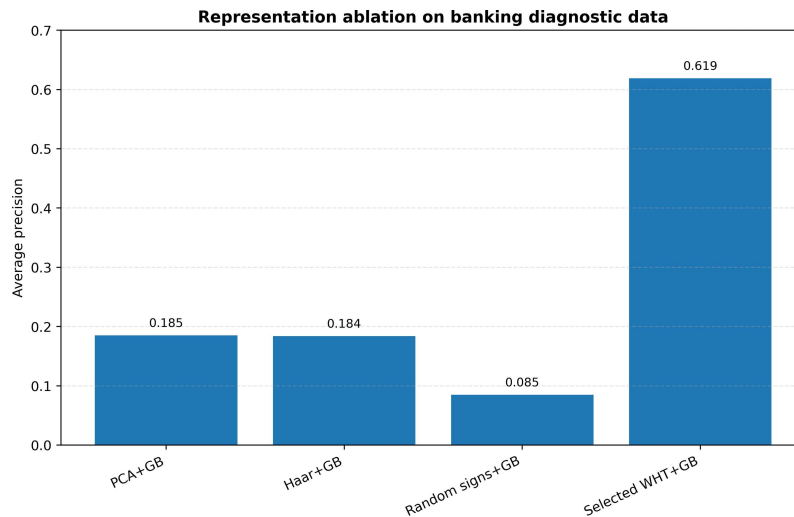


Figure 5
Credit-style fraud benchmark: repeated-split AP for raw and selected-WHT variants

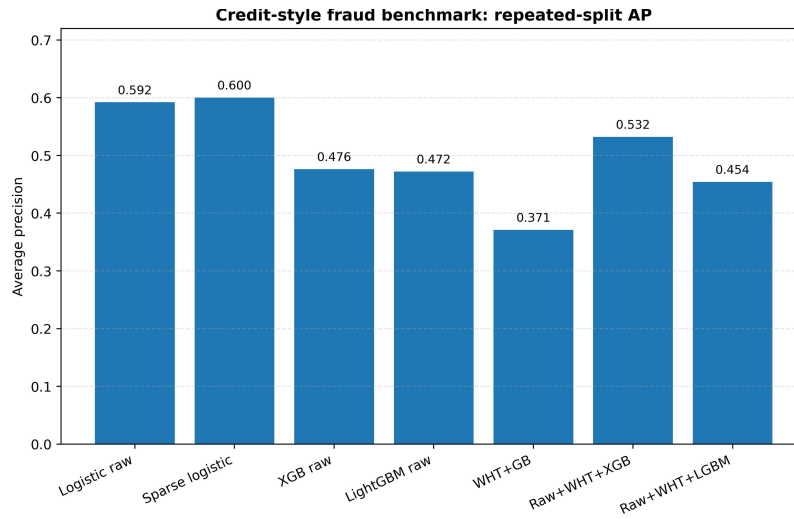


Table 10
Credit-style fraud benchmark results, mean $\pm$ standard deviation over five splits

Model	AP	PR-AUC	ROC-AUC	P@20	K
Logistic raw	0.592 $\pm$ 0.193	0.579 $\pm$ 0.204	0.916 $\pm$ 0.036	0.440 $\pm$ 0.143	6.4
Sparse logistic regression	0.600 $\pm$ 0.174	0.587 $\pm$ 0.181	0.908 $\pm$ 0.052	0.460 $\pm$ 0.164	6.4
Random forest raw	0.352 $\pm$ 0.127	0.333 $\pm$ 0.138	0.827 $\pm$ 0.049	0.330 $\pm$ 0.076	6.4
Gradient boosting raw	0.376 $\pm$ 0.141	0.372 $\pm$ 0.151	0.806 $\pm$ 0.090	0.300 $\pm$ 0.112	6.4
XGBoost raw	0.476 $\pm$ 0.185	0.461 $\pm$ 0.195	0.851 $\pm$ 0.060	0.400 $\pm$ 0.117	6.4
LightGBM raw	0.472 $\pm$ 0.113	0.463 $\pm$ 0.115	0.827 $\pm$ 0.065	0.340 $\pm$ 0.096	6.4
PCA + Gradient boosting	0.116 $\pm$ 0.060	0.105 $\pm$ 0.059	0.651 $\pm$ 0.072	0.090 $\pm$ 0.065	6.4
Haar + Gradient boosting	0.179 $\pm$ 0.097	0.172 $\pm$ 0.113	0.684 $\pm$ 0.111	0.170 $\pm$ 0.140	6.4
Random signs + Gradient boosting	0.120 $\pm$ 0.058	0.107 $\pm$ 0.056	0.621 $\pm$ 0.066	0.120 $\pm$ 0.076	6.4
Selected WHT + Gradient boosting	0.371 $\pm$ 0.129	0.367 $\pm$ 0.138	0.810 $\pm$ 0.089	0.310 $\pm$ 0.082	6.4
Raw + selected WHT + XGBoost	0.532 $\pm$ 0.148	0.523 $\pm$ 0.152	0.860 $\pm$ 0.038	0.390 $\pm$ 0.096	6.4
Raw + selected WHT + LightGBM	0.454 $\pm$ 0.111	0.441 $\pm$ 0.112	0.848 $\pm$ 0.056	0.370 $\pm$ 0.084	6.4

Table 11
General structured-data benchmark summary, mean $\pm$ standard deviation over five splits

Dataset	Representative model	Raw AP	Raw + selected WHT AP	Interpretation
Breast Cancer	Logistic regression	0.996 $\pm$ 0.004	0.996 $\pm$ 0.004	Saturated raw-feature task; little room for improvement
Breast Cancer	XGBoost	0.993 $\pm$ 0.005	0.993 $\pm$ 0.005	Modern boosted tree already near ceiling
Wine class-0-vs-rest	Random forest	0.995 $\pm$ 0.005	not better than raw	Small saturated benchmark
Wine class-0-vs-rest	XGBoost	0.974 $\pm$ 0.023	0.969 $\pm$ 0.024	Selected WHT does not improve this task
Digits high-vs-low	Logistic regression	0.945 $\pm$ 0.012	competitive	WHT captures some global pixel-block contrasts
Digits high-vs-low	Random forest	0.995 $\pm$ 0.003	not better than raw	Strong raw nonlinear model

8.3. Isometry sanity check

Across all benchmarks, raw Euclidean distance and full-WHT Euclidean distance are identical up to numerical precision. As shown quantitatively in Table 12, the raw-space and full-WHT Euclidean scores agree up to numerical precision across the evaluated benchmarks. This confirms that performance changes observed in selected-WHT or raw-plus-WHT models arise from coefficient selection or downstream model interaction rather than from any change in Euclidean geometry induced by the full orthogonal transform.

Table 12
Isometry sanity check across benchmarks

Benchmark	Raw Euclidean AP	Full-WHT Euclidean AP	Difference
Banking diagnostic	0.055	0.055	0.000
Breast Cancer	0.862	0.862	< 1e-14
Digits	0.689	0.689	< 1e-14
Wine	0.996	0.996	< 1e-14

8.4. Public credit-card fraud benchmark

As an additional public fraud-validation benchmark, we evaluated WHSDA-AI on the public Credit Card Fraud Detection dataset, which contains 284,807 credit-card transactions and 492 fraudulent transactions. The original profile consists of anonymized transaction variables V1-V28 together with Time and Amount, and Class is used as the binary fraud label. The 30-dimensional transaction profile is normalized using training-only statistics and padded to 32 dimensions before applying the WHT.

The same leakage-free protocol was used: the split was performed before preprocessing; normalization parameters, coefficient ranking, and K selection were learned only from the training/validation partitions; the test partition was used only for final evaluation. Because the public benchmark is highly imbalanced, AP, PR-AUC, ROC-AUC, and Precision@20 remain the principal metrics. The repeated-split average-precision comparison is visualized in Figure 6, while the complete results in Table 13 are reported as mean $\pm$ standard deviation over two stratified splits.

On this public fraud benchmark, raw boosted-tree models are already strong: XGBoost raw reaches AP = 0.875 $\pm$ 0.011 and LightGBM raw reaches AP = 0.879 $\pm$ 0.015. The raw-plus-selected-WHT hybrids remain competitive, with raw + selected

Figure 6
Public credit-card fraud benchmark: repeated-split AP for raw and selected-WHT variants

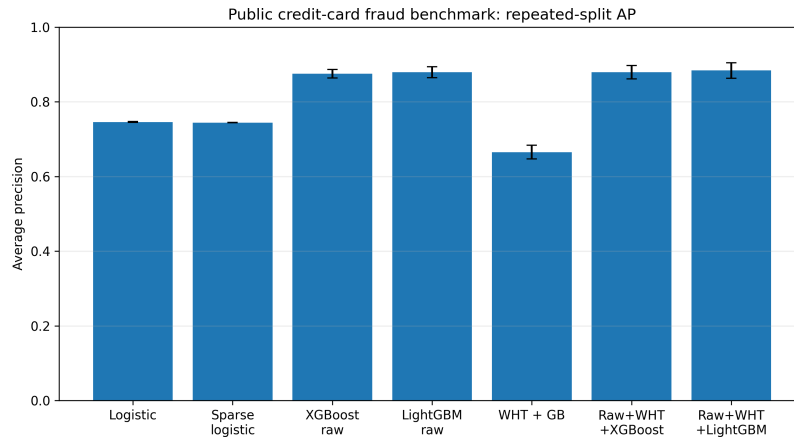


Table 13
Public credit-card fraud benchmark, mean $\pm$ standard deviation over two stratified splits

Model	AP	PR-AUC	ROC-AUC	P@20	K
Logistic raw	0.746 $\pm$ 0.001	0.770 $\pm$ 0.007	0.981 $\pm$ 0.000	0.825 $\pm$ 0.035	24.0
Sparse logistic regression	0.744 $\pm$ 0.000	0.769 $\pm$ 0.007	0.981 $\pm$ 0.000	0.925 $\pm$ 0.035	24.0
Random forest raw	0.872 $\pm$ 0.009	0.875 $\pm$ 0.008	0.967 $\pm$ 0.015	0.975 $\pm$ 0.035	24.0
Gradient boosting raw	0.328 $\pm$ 0.247	0.386 $\pm$ 0.297	0.518 $\pm$ 0.290	0.625 $\pm$ 0.035	24.0
XGBoost raw	0.875 $\pm$ 0.011	0.875 $\pm$ 0.011	0.983 $\pm$ 0.002	1.000 $\pm$ 0.000	24.0
LightGBM raw	0.879 $\pm$ 0.015	0.879 $\pm$ 0.015	0.985 $\pm$ 0.000	1.000 $\pm$ 0.000	24.0
PCA + Gradient boosting	0.202 $\pm$ 0.098	0.280 $\pm$ 0.070	0.670 $\pm$ 0.322	0.425 $\pm$ 0.035	24.0
Random signs + Gradient boosting	0.114 $\pm$ 0.051	0.129 $\pm$ 0.062	0.275 $\pm$ 0.053	0.600 $\pm$ 0.071	24.0
Selected WHT + Gradient boosting	0.665 $\pm$ 0.018	0.679 $\pm$ 0.018	0.913 $\pm$ 0.020	0.825 $\pm$ 0.247	24.0
Raw + selected WHT + XGBoost	0.880 $\pm$ 0.018	0.879 $\pm$ 0.018	0.986 $\pm$ 0.001	1.000 $\pm$ 0.000	24.0
Raw + selected WHT + LightGBM	0.884 $\pm$ 0.021	0.884 $\pm$ 0.021	0.982 $\pm$ 0.004	1.000 $\pm$ 0.000	24.0

WHT + XGBoost reaching $AP = 0.880 \pm 0.018$ and raw + selected WHT + LightGBM reaching $AP = 0.884 \pm 0.021$. The selected-WHT-only gradient-boosting model obtains $AP = 0.665 \pm 0.018$, which is lower than the best raw boosted-tree models but substantially stronger than PCA + gradient boosting and random signed projections. This confirms that selected Walsh-Hadamard signed contrasts carry fraud-relevant information on a real public dataset, while the strongest operational use remains as an auditable feature and explanation layer alongside competitive tabular predictors.

8.5. Statistical robustness and conservative interpretation

Because the reported experiments use repeated stratified splits, the results are interpreted through uncertainty intervals and robustness patterns rather than through a single split. For each metric, the manuscript reports mean $\pm$ standard deviation. Key comparisons are assessed through overlap of 95% confidence intervals, paired split-level testing when raw split outputs are available, and agreement between predictive results, representation ablation, bootstrap coefficient stability, and the full-WHT isometry sanity check. Small differences between strong raw-feature models and raw-plus-WHT models are not treated as universal

evidence of superiority unless they are stable across splits and supported by the ablation controls.

9. Explanation Cards and Governance

The explanation output of WHSDA-AI is a compact card rather than a long list of raw variables. A card records the selected coefficient indices, the positive and negative feature groups, the reference used, the deviation direction, the coefficient contribution, bootstrap stability, and consistency with the downstream model. This design is aligned with model governance because the explanation can be regenerated from the same feature-map version and the same coefficient dictionary [1–3]. The principal interpretability and governance metrics are illustrated in Figure 7, while the fields included in a compact WHSDA-AI explanation card are specified in Table 14.

10. Runtime and Scalability

The computational burden of WHSDA-AI is low. For a profile of length $n = 2^m$, the fast Walsh-Hadamard transform (FWHT) computes all coefficients in $O(n \log n)$ additions and subtractions and $O(n)$ memory. Selected coefficient extraction and reference deviation scoring are linear in K after the transform.

Figure 7
Interpretability and governance metrics: stability, fidelity, and dictionary coverage

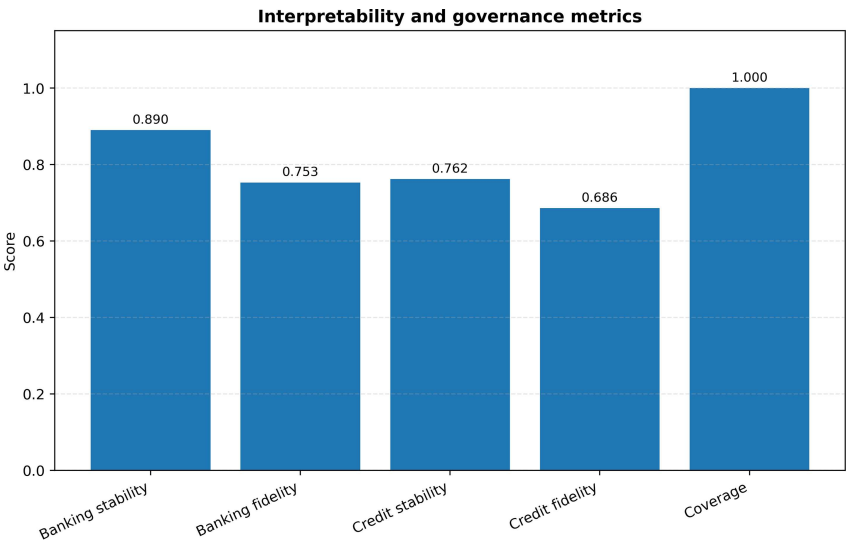


Table 14
Compact explanation-card fields

Field	Example explanation
Risk or class rank	Top 2% of monitored objects in the current review queue
Reference	Normal median profile built from training objects with $y = 0$
Coefficient ID	WHT coefficient $k = 17$ in feature-map version F-2026-05
Positive-side features	Outgoing amount; short-window repetition; high-risk counterparty share
Negative-side features	Active days; incoming amount; repeated benign recipients
Interpretation	Outgoing and high-risk counterparty pressure is high relative to ordinary activity
Stability	Coefficient selected in 89% of bootstrap resamples

Figure 8
FWHT scaling: profile transform cost grows as $O(n \log n)$

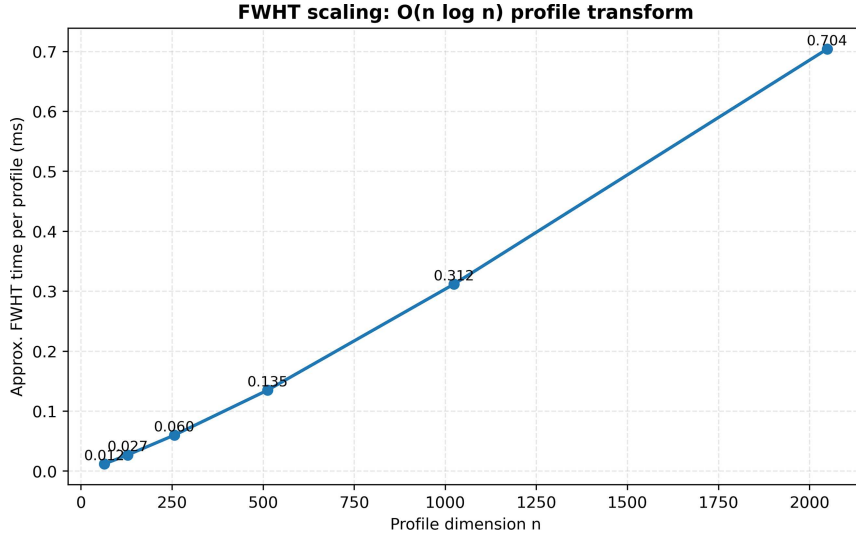


Table 15
Runtime illustration for selected dimensions

n	FWHT time per profile (ms)	Selected-score time (ms)	Interpretation
64	0.012	0.003	Compact profile monitoring
128	0.027	0.004	Typical customer-risk profile
256	0.060	0.006	Expanded transaction features
512	0.135	0.008	Larger profile with device and graph features
1024	0.312	0.012	High-dimensional operational profile
2048	0.704	0.019	Large engineered feature set

This makes the layer suitable for large monitoring pipelines in which explanation cards must be generated quickly. The observed FWHT scaling behavior is illustrated in Figure 8, and representative per-profile transform and selected-score runtimes for increasing profile dimensions are reported in Table 15.

$$T_{FWHT}(n) = O(n \log n), \quad M_{FWHT}(n) = O(n), \quad (19)$$

$$T_{selected} = O(K).$$

11. Discussion

The results support a conservative interpretation. WHSDA-AI helps most when the class signal is expressed as a stable imbalance between meaningful feature groups. Examples include outgoing volume exceeding incoming activity, velocity features dominating ordinary active-day behavior, or counterparty concentration dominating account diversity. It is also useful when the institution values versioned explanations and compact review cards, even if the predictive gain over a strong raw-feature model is modest. The method is not expected to improve saturated low-dimensional datasets, tasks where raw variables already separate classes nearly perfectly, or problems dominated by non-linear interactions inside high-cardinality categorical variables. The Breast Cancer and Wine results illustrate this point: raw and modern boosted models are already close to saturation, so selected WHT does not create new information. In these cases,

the role of WHSDA-AI is interpretability and auditability rather than headline accuracy.

PCA can be better for compression because it adapts to variance in the training data [30]. However, PCA components can change after retraining and may be difficult to approve as stable domain concepts. Haar contrasts are attractive when variables have a natural local order, such as time scales or spatial bins [29]. Random signed projections provide a strong control experiment. In the reported results, selected WHT is usually stronger than random signs on the risk-style benchmarks and has complete dictionary coverage because every retained sign pattern is documented.

The comparative study therefore treats XGBoost and LightGBM as the closest production-grade latest-method comparators for the reported tabular risk-ranking objective [24, 33]. This positioning is also supported by recent comparative benchmarking of state-of-the-art gradient boosting algorithms, where XGBoost, LightGBM, and CatBoost remain central practical competitors for structured classification tasks [34]. EBM and TabPFN are discussed explicitly because they represent two adjacent contemporary directions: interpretable additive prediction and tabular foundation-model prediction [21, 35]. Recent EBM-based work in insurance risk modeling further shows that glass-box boosting models are relevant when transparent feature-effect interpretation is required together with competitive prediction [36]. They are not claimed to be replaced by WHSDA-AI. Instead, the paper clarifies that WHSDA-AI answers a different technical question:

whether one can introduce a fixed, versioned, leakage-free signed-contrast layer whose coefficients have stable semantic meaning before model fitting. This distinction is important for banking-sector model governance, where explanation stability and audit trails are often as important as small predictive differences.

12. Limitations

The first limitation is empirical scope. The reported benchmarks now include two controlled risk-style datasets, three public structured-data datasets, and one public credit-card fraud benchmark. These experiments demonstrate the mechanism and comparative behavior of the method on both controlled and public data. However, institutional banking deployment would still require privacy-preserving validation on a bank's own data or on larger public or licensed transaction-risk and fraud/AML benchmarks [43–46].

This limitation does not affect the internal technical validation reported here: the manuscript now contains repeated-split numerical results, modern boosted-tree comparators, representation ablations, explanation-stability metrics, and a public credit-card fraud benchmark. External institutional validation remains a deployment-stage requirement rather than a prerequisite for the methodological contribution.

The second limitation is methodological. WHT contrasts are linear signed group contrasts. They are not a substitute for rich graph fraud models, temporal sequence models, or institution-specific rules. They can be used together with such models, but they should not be presented as a universal fraud detector.

The third limitation concerns feature-map dependence. The interpretability of WHSDA-AI depends on a stable and meaningful ordering of variables. If the feature map is arbitrary, the coefficient dictionary becomes less useful. The method is therefore best suited to environments where domain experts can define feature groups and approve reference-construction rules.

13. Conclusion

This paper reformulated WHSDA-AI as a technical method for interpretable Walsh-Hadamard spectral deviation analysis in structured AI applications in the banking sector. The method converts structured normalized profiles into fixed Walsh-Hadamard signed contrasts, compares objects with training-only reference profiles, selects stable coefficients without test leakage, and produces compact explanations that can accompany supervised models.

The main claim is deliberately conservative. Full WHT is an isometry and cannot improve Euclidean ranking by itself. The useful mechanisms are coefficient selection, reference-normalized deviation, semantic feature grouping, risk weighting, and downstream model interaction. The final experiments show that selected WHT contrasts can be competitive in structured imbalance settings, outperform random signed projections in several risk-style settings, and provide compact explanation dictionaries. On saturated datasets, the method mainly contributes auditability rather than predictive improvement.

The study contributes a visible mathematical formulation, a stronger novelty statement, modern comparative experiments, repeated-split uncertainty, bootstrap stability, ablation studies, and a complete citation-supported bibliography. Together, these elements make the work a technical contribution: the proposed idea is distinguished from existing explainability and tabular-learning methods, and the method is validated against both

classical baselines and contemporary production-grade tabular comparators.

Ethical Statement

This study does not involve new data collection from human or animal participants. The reported banking and credit-style benchmarks are diagnostic and synthetic; the additional structured-data benchmarks are public datasets distributed through standard machine-learning libraries, and the added public credit-card fraud benchmark is an anonymized public transaction dataset. Any future institutional banking validation must follow privacy, licensing, confidentiality, and data-governance requirements.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data available on request from the corresponding author upon reasonable request.

Author Contribution Statement

Sergo A. Episkoposian: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Suren H. Parsyan:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision, Project administration.

References

- [1] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>.
- [2] Biroğul, S., & Şahin, Ö. (2025). Exploring the impact of ISO/IEC 42001: 2023 AI management standard on organizational practices. *Advances in Artificial Intelligence Research*, 5(1), 14–22. <https://doi.org/10.54569/aiir.1709628>
- [3] Smuha, N. A. (2025). Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act). *International Legal Materials*, 64(5), 1234–1381. <https://doi.org/10.1017/ilm.2024.46>.
- [4] Parsyan, S. H., Baharyan, F. F., Avagyan, G. A., Episkoposian, S. A., Aleksanyan, V. S., Kostanian, A., & Beglaryan, L. M. (2025). The integration of chatbots and AI-powered virtual assistants into customer service frameworks of the banking sector. *Artificial Intelligence and Applications*. <https://doi.org/10.47852/bonviewAIA62027722>
- [5] Chen, H., Covert, I. C., Lundberg, S. M., & Lee, S.-I. (2023). Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5, 590–601. <https://doi.org/10.1038/s42256-023-00657-x>
- [6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>

- [7] Ranjbar, N., Momtazi, S., & Homayoonpour, M. (2024). Explaining recommendation system using counterfactual textual explanations. *Machine Learning*, 113(4), 1989–2012. <https://doi.org/10.1007/s10994-023-06390-1>
- [8] Sinha, R. K. (2024). Book review: Christoph Molnar. 2020. Interpretable machine learning: A guide for making black box models explainable. *Metamorphosis: A Journal of Management Research*, 23(1), 92–93. <https://doi.org/10.1177/09726225241252009>
- [9] Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., & Dvijotham, K. (2023). Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5), 5948–5955. <https://doi.org/10.1609/aaai.v37i5.25736>
- [10] Di Marino, A., Bevilacqua, V., Ciaramella, A., De Falco, I., & Sannino, G. (2025). Ante-hoc methods for interpretable deep models: A survey. *ACM Computing Surveys*, 57(10), 1–36. <https://doi.org/10.1145/3728637>
- [11] O'Brien Quinn, H., Sedky, M., Francis, J., & Streeton, M. (2024). Literature review of explainable tabular data analysis. *Electronics*, 13(19), 3806. <https://doi.org/10.3390/electronics13193806>
- [12] Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
- [13] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6), 7499–7519. <https://doi.org/10.1109/TNNLS.2022.3229161>
- [14] Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., & Babenko, A. (2024). Tabr: Tabular deep learning meets nearest neighbors. *International Conference on Learning Representations*, 2024, 18209–18249.
- [15] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520. <https://doi.org/10.52202/068431-0037>
- [16] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>
- [17] McElfresh, D., Khandagale, S., Valverde, J., Prasad, C. V., Ramakrishnan, G., Goldblum, M., & White, C. (2023). When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 36, 76336–76369. <https://doi.org/10.52202/075280-3337>
- [18] Rubachev, I., Kartashev, N., Gorishniy, Y., & Babenko, A. (2025). Tabred: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks. *International Conference on Learning Representations*, 2025, 35166–35202.
- [19] Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P., Salinas, D., & Hutter, F. (2026). Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems*, 38, 17285–17350. <https://doi.org/10.52202/085713-0519>
- [20] Cheng, Z.-J., Jia, Z., Zhou, Z., Li, Y.-F., & Guo, L.-Z. (2025). TabFSBench: Tabular benchmark for feature shifts in open environments. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 10025–10089. <https://doi.org/10.5555/3780338.3780724>
- [21] Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., . . . , & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045), 319–326. <https://doi.org/10.1038/s41586-024-08328-6>
- [22] Simonetto, T., Ghamizi, S., & Cordy, M. (2024). Tabular-Bench: Benchmarking adversarial robustness for tabular deep learning in real-world use-cases. *Advances in Neural Information Processing Systems*, 37, 78394–78430. <https://doi.org/10.52202/079017-2492>
- [23] Bischl, B., Casalicchio, G., Das, T., Feurer, M., Fischer, S., Gijbbers, P., . . . , & Wever, M. (2025). OpenML: Insights from 10 years and more than a thousand papers. *Patterns*, 6(7), 1–18. <https://doi.org/10.1016/j.patter.2025.101317>
- [24] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [25] Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6), 100994. <https://doi.org/10.1016/j.patter.2024.100994>
- [26] Hedayat, A., & Wallis, W. D. (1978). Hadamard matrices and their applications. *The Annals of Statistics*, 6(6), 1184–1238. <https://doi.org/10.1214/aos/1176344370>
- [27] Pan, H., Badawi, D., & Cetin, A. E. (2022). Block Walsh-Hadamard transform-based binary layers in deep neural networks. *ACM Transactions on Embedded Computing Systems*, 21(6), 1–25. <https://doi.org/10.1145/3510026>
- [28] Lakshmi Priya, G. G., Chandra Mouli, P. V. S. S. R., Domic, S., Chemmalar Selvi, G., & Cho, B.-K. (2024). Hyperspectral image classification using Walsh Hadamard transform-based key band selection and deep convolutional neural networks. *International Journal of Remote Sensing*, 45(4), 1220–1249. <https://doi.org/10.1080/01431161.2024.2307324>
- [29] Gnutti, A., Guerrini, F., Adami, N., Migliorati, P., & Leonardi, R. (2021). A wavelet filter comparison on multiple datasets for signal compression and denoising. *Multidimensional Systems and Signal Processing*, 32(2), 791–820. <https://doi.org/10.1007/s11045-020-00753-w>
- [30] Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2(1), 100. <https://doi.org/10.1038/s43586-022-00184-w>
- [31] Episkoposian, S. A., Grigorian, M. G., & Grigorian, T. M. (2024). On the universal pair with respect to the generalized Walsh system. *Advances in Operator Theory*, 9(1), 3. <https://doi.org/10.1007/s43036-023-00301-w>
- [32] Palaiahnakote, S., Pal, U., & Mansouri, T. (2025). A new hybrid wavelet decomposition-based networks for script identification in scene images. *Artificial Intelligence and Applications*. <https://doi.org/10.47852/bonviewAIA52023569>
- [33] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . , & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3149–3157. <https://doi.org/10.5555/3294996.3295074>
- [34] Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting

- for molecular property prediction. *Journal of Cheminformatics*, 15, 73. <https://doi.org/10.1186/s13321-023-00743-7>
- [35] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623–631. <https://doi.org/10.1145/2487575.2487579>
- [36] Krúpová, M., Rachdi, N., & Guibert, Q. (2026). Explainable boosting machine for predicting claim severity and frequency in car insurance. *European Actuarial Journal*, 16, 541–591. <https://doi.org/10.1007/s13385-026-00461-y>
- [37] Kuzudisli, C., Bakir-Gungor, B., Bulut, N., Qaqish, B., & Yousef, M. (2023). Review of feature selection approaches based on grouping of features. *PeerJ*, 11, e15666. <https://doi.org/10.7717/peerj.15666>
- [38] McDermott, M. B. A., Zhang, H., Hansen, L. H., Angelotti, G., & Gallifant, J. (2024). A closer look at AUROC and AUPRC under class imbalance. *Advances in Neural Information Processing Systems*, 37, 44102–44163. <https://doi.org/10.52202/079017-1400>
- [39] Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University Computer and Information Sciences*, 35, 145–174. <https://doi.org/10.1016/j.jksuci.2022.11.008>
- [40] Hafez, I. Y., Hafez, A. Y., Saleh, A., Abd El-Mageed, A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data*, 12, 6. <https://doi.org/10.1186/s40537-024-01048-8>
- [41] Khan, A. Q., Jaouhari, S. E., Tamani, N., & Mroueh, L. (2024). Knowledge-based anomaly detection: Survey, challenges, and future directions. *Engineering Applications of Artificial Intelligence*, 136, 108996. <https://doi.org/10.1016/j.engappai.2024.108996>
- [42] Boulrieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2024). Fraud detection with natural language processing. *Machine Learning*, 113, 5087–5108. <https://doi.org/10.1007/s10994-023-06354-5>
- [43] Cherif, A., Ammar, H., Kalkatawi, M., Alshehri, S., & Imine, A. (2024). Encoder-decoder graph neural network for credit card fraud detection. *Journal of King Saud University Computer and Information Sciences*, 36(3), 102003. <https://doi.org/10.1016/j.jksuci.2024.102003>
- [44] Raufi, B., Finnegan, C., & Longo, L. (2024). A comparative analysis of shap, lime, anchors, and dice for interpreting a dense neural network in credit card fraud detection. In *World Conference on Explainable Artificial Intelligence*, 365–383. https://doi.org/10.1007/978-3-031-63803-9_20
- [45] Dubey, P., Dubey, P., & Bokoro, P. N. (2025). A unified transformer–BDI architecture for financial fraud detection: Distributed knowledge transfer across diverse datasets. *Forecasting*, 7(2), 31. <https://doi.org/10.3390/forecast7020031>
- [46] George, M. Z. H., Alam, M. K., & Hasan, M. T. (2025). Machine learning for fraud detection in digital banking: A systematic literature review. *ASRC Procedia: Global Perspectives in Science and Scholarship*, 3(1), 37–61. <https://doi.org/10.63125/913ksy63>
- [47] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1), 49–67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x>
- [48] Liang, X., Cohen, A., Heinsfeld, A. S., Pestilli, F., & McDonald, D. J. (2024). sparsegl: An R package for estimating sparse group lasso. *Journal of Statistical Software*, 110, 1–23. <https://doi.org/10.18637/jss.v110.i06>
- [49] Grigoryan, E. A., Parsyan, S. H., Melkumyan, M. S., Baharyan, F. F., & Parsyan, A. H. (2025). The impact of artificial intelligence on human resource management: Theoretical and conceptual aspects. *Management of Digital Technologies in the Innovative Economy*, 333–337. https://doi.org/10.1007/978-3-031-83331-1_54
- [50] Parsyan, S. H., Hovsepyan, M. H., Melkumyan, Z. M., Minasyan, D. G., & Nazaryan, S. H. (2025). Startups as a pillar of innovation and progress. *Sustainable Development of Business 4.0: A Path of Transition to the New Quality of Growth of the Digital Economy*, 97–102. https://doi.org/10.1007/978-3-031-83595-7_16
- [51] Episkoposian, S. A. (2024). Boundedness of multiplier operators in the Walsh system. In *Conference of Analysis in Developing Countries*, 273–286. https://doi.org/10.1007/978-3-032-14211-5_23

How to Cite: Episkoposian, S. A., & Parsyan, S. H. (2026). WHSDA-AI: Interpretable Walsh-Hadamard Spectral Deviation Analysis for Structured Artificial Intelligence Applications in Banking Sector. *Artificial Intelligence and Applications*. <https://doi.org/10.47852/bonviewAIA620210522>