CWAE-MMD: Conditional Wasserstein Autoencoders for Contextual Anomaly Detection in Transaction Monitoring

Anonymous Author(s)Anonymous Institution

***Abstract*—Contextual anomaly detection—identifying samples whose behavioral features deviate from expectations given their context—is critical for anti-money laundering (AML) systems. Conditional Variational Autoencoders (CVAEs) offer a principled approach by learning context-conditioned reconstructions, but suffer from numerical instability when applied to highly skewed tabular data common in financial applications: the KL divergence regularization requires per-sample variance prediction, which can diverge and produce NaN values. We study CWAE-MMD (Conditional Wasserstein Autoencoder with Maximum Mean Discrepancy), which replaces per-sample KL with aggregate posterior matching via MMD. This enables deterministic encoding without variance prediction, eliminating the primary source of numerical instability while allowing the latent space to form distinct context-dependent clusters. For anomaly scoring, we employ a two-stage pipeline: normalized reconstruction residuals are fed to Isolation Forest, providing non-parametric anomaly scores without assuming a specific error distribution. We systematically evaluate nine CWAE-MMD configurations—spanning scoring strategies, reconstruction losses, and kernel settings—against CVAE, Isolation Forest, and Deep Isolation Forest baselines across 15 datasets. On contextual synthetic benchmarks, CWAE-MMD achieves up to 1.000 AUROC, outperforming IF by 19 points. On real-world benchmarks including PaySim fraud data, CWAE-MMD variants outperform all baselines, achieving 0.913 AUROC where standard IF reaches 0.789.**

***Index Terms*—contextual anomaly detection, Wasserstein autoencoder, MMD, anti-money laundering, isolation forest**

I. Introduction

Contextual anomaly detection identifies samples whose behavioral features deviate from what is expected given their context. In transaction monitoring, a $500,000 wire transfer may be routine for a corporate treasury account but highly suspicious for a retail checking account. The challenge lies in modeling the context-behavior relationship so that deviations can be measured against the right reference distribution [1], [2].

Conditional Variational Autoencoders (CVAEs) [3], [4] offer a principled approach to this problem by learning a generative model $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})$ conditioned on context. However, CVAEs suffer from significant practical limitations when applied to real-world tabular data. The KL divergence regularization term, which enforces each sample's posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})$ to match the prior $p(\mathbf{z})$, becomes numerically unstable with highly skewed feature distributions common in financial data. The encoder must predict both mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$ for the latent distribution, and extreme feature values can cause $\log \boldsymbol{\sigma}^2$ to explode, resulting in NaN values during training. Furthermore, the per-sample KL constraint forces the latent space to be a "fuzzy" Gaussian everywhere, potentially causing the model to reconstruct anomalies too well by learning an overly smooth representation.

Wasserstein Autoencoders (WAEs) [5] address these stability issues by replacing the per-sample KL divergence with aggregate posterior matching via Maximum Mean Discrepancy (MMD) [6]. Rather than constraining each $q(\mathbf{z}|\mathbf{x})$ individually, WAE enforces that the aggregate posterior $Q_Z = \mathbb{E}_{p(\mathbf{x})}[q(\mathbf{z}|\mathbf{x})]$ matches the prior P_Z. This formulation allows deterministic encoders (no variance prediction), eliminates sampling noise, and permits the latent space to form distinct clusters while maintaining global distributional properties. However, the original WAE is unconditional and not directly applicable to contextual anomaly detection.

In this paper, we study CWAE-MMD (Conditional Wasserstein Autoencoder with Maximum Mean Discrepancy), which combines the contextual modeling capability of CVAE with the numerical stability of WAE. The approach conditions both encoder and decoder on context features $\mathbf{c}$, uses MMD to match the aggregate posterior to a Gaussian prior, and employs a deterministic encoder for reproducible inference. For anomaly scoring, we employ a two-stage pipeline: normalized reconstruction residuals are fed to an Isolation Forest [7], providing non-parametric anomaly scores without assuming a specific error distribution.

We systematically evaluate multiple CWAE-MMD configurations and make the following contributions:

- We combine conditional autoencoding with aggregate posterior matching via MMD, inheriting the stability of WAE while enabling contextual anomaly detection.
- We study a two-stage anomaly scoring pipeline where Isolation Forest operates on normalized reconstruction residuals, decoupling the reconstruction model from the scoring mechanism.
- We evaluate nine CWAE-MMD configurations—spanning scoring strategies (IF, DIF, latent, dual-space), reconstruction losses (MSE, CRPS, Huber), and kernel settings (standard IMQ, IMQ-Extreme)—identifying when each variant is most effective.
- We provide experimental evaluation across 15 datasets comparing CWAE-MMD against CVAE, Isolation Forest,

and Deep Isolation Forest baselines.

The remainder of this paper is organized as follows. Section II presents the CWAE-MMD method, including the base architecture, all scoring and loss variants, and the relationship to prior work. Section III describes experimental results. Section IV discusses findings and concludes.

II. CWAE-MMD Method

A. Background: From CVAE to WAE

CVAEs [3], [8] learn a stochastic encoder $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})$ and decoder $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{c})$, optimizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{c})] - \beta \cdot \text{KL}[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})\|p(\mathbf{z})] \quad (1)$$

The KL term requires per-sample distribution parameters $\boldsymbol{\mu}(\mathbf{x},\mathbf{c})$ and $\log \boldsymbol{\sigma}^2(\mathbf{x},\mathbf{c})$:

$$\text{KL} = -\tfrac{1}{2}\textstyle\sum_j \big(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\big) \quad (2)$$

This formulation has three instability sources: (1) $\log\sigma^2 \to -\infty$ when the encoder predicts near-zero variance; (2) $\log\sigma^2 \to +\infty$ causes $\exp(\log\sigma^2)$ to overflow; (3) the reparameterization trick $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ amplifies noise when $\boldsymbol{\sigma}$ is extreme.

Wasserstein Autoencoders [5] replace per-sample KL with aggregate posterior matching. The WAE objective is:

$$\mathcal{L}_{\text{WAE}} = \mathbb{E}_{p(\mathbf{x})}[c(\mathbf{x}, G(E(\mathbf{x})))] + \lambda \cdot D_Z(Q_Z, P_Z) \quad (3)$$

where E is the encoder, G is the decoder, $c(\cdot,\cdot)$ is a reconstruction cost, and D_Z measures divergence between the aggregate posterior and prior. The WAE-MMD variant uses Maximum Mean Discrepancy:

$$\text{MMD}^2(P,Q) = \mathbb{E}[k(\mathbf{z},\mathbf{z}')] - 2\mathbb{E}[k(\mathbf{z},\tilde{\mathbf{z}})] + \mathbb{E}[k(\tilde{\mathbf{z}},\tilde{\mathbf{z}}')] \quad (4)$$

where $\mathbf{z},\mathbf{z}' \sim P_Z$, $\tilde{\mathbf{z}},\tilde{\mathbf{z}}' \sim Q_Z$, and $k(\cdot,\cdot)$ is a kernel function. The inverse multiquadrics (IMQ) kernel $k(\mathbf{x},\mathbf{y}) = C/(C + \|\mathbf{x}-\mathbf{y}\|^2)$ is preferred for its heavier tails compared to RBF, preventing vanishing gradients for distant points.

WAE offers key advantages: (1) deterministic encoders are permitted since no reparameterization trick is needed, (2) no variance prediction eliminates $\log/\exp$ numerical instability, and (3) the aggregate constraint allows latent codes to form distinct clusters while maintaining global Gaussian statistics. However, the original WAE is unconditional.

B. CWAE-MMD Architecture

CWAE-MMD extends WAE to the conditional setting. The encoder maps behavioral features and context to a latent code:

$$\mathbf{z} = E_\phi(\mathbf{x},\mathbf{c}) = \text{MLP}_\phi([\mathbf{x};\mathbf{c}]) \quad (5)$$

where $[\mathbf{x};\mathbf{c}]$ denotes concatenation and $\mathbf{z} \in \mathbb{R}^{d_z}$. Critically, the encoder is deterministic—it outputs a point estimate rather than distribution parameters. The decoder reconstructs behavioral features:

$$\hat{\mathbf{x}} = D_\theta(\mathbf{z},\mathbf{c}) = \text{MLP}_\theta([\mathbf{z};\mathbf{c}]) \quad (6)$$

Both encoder and decoder are multi-layer perceptrons with hidden layers $[128, 64]$ and ReLU activations. Context $\mathbf{c}$ conditions both networks, allowing the model to learn context-dependent reconstructions.

The training objective combines reconstruction loss with MMD regularization:

$$\mathcal{L} = \mathcal{L}_{\text{recon}}(\mathbf{x},\hat{\mathbf{x}}) + \lambda \cdot \text{MMD}^2(Q_Z, P_Z) \quad (7)$$

We use multi-scale IMQ kernels with bandwidth $C_s = s \cdot 2d_z$ at scales $s \in \{0.2, 0.5, 1.0, 2.0, 5.0\}$. In practice, MMD is estimated on mini-batches of size B:

$$\widehat{\text{MMD}}^2 = \frac{1}{B(B-1)}\sum_{i\neq j} k(\mathbf{z}_i,\mathbf{z}_j) - \frac{2}{B^2}\sum_{i,j} k(\mathbf{z}_i,\tilde{\mathbf{z}}_j) + \frac{1}{B(B-1)}\sum_{i\neq j} k(\tilde{\mathbf{z}}_i,\tilde{\mathbf{z}}_j) \quad (8)$$

Unlike KL divergence where $\log\sigma^2$ can diverge, the IMQ kernel is bounded in $(0,1]$, ensuring $\text{MMD}^2 \in [0,2]$ with bounded gradients. This eliminates the need for log-variance clamping commonly required in CVAE implementations. We optimize using Adam with learning rate 10^{-3} and weight decay 10^{-5}, with early stopping on validation loss.

C. Scoring Variants

After training, anomaly scores are derived from reconstruction residuals $\mathbf{r} = \mathbf{x} - \hat{\mathbf{x}}$, normalized by training residual standard deviation: $\mathbf{r}_{\text{norm}} = \mathbf{r}/\boldsymbol{\sigma}_{\text{train}}$.

CWAE-IF (base scorer). Fits an Isolation Forest [7] on $\mathbf{r}_{\text{norm}}$: $s_i = -\text{IF}(\mathbf{r}_{\text{norm},i})$. *Pro:* Simple, fast, well-understood. *Con:* Axis-aligned splits may miss non-linear structure in residual space.

CWAE-DIF (Deep Isolation Forest [9]). Applies M=6 frozen random MLP projections ϕ_m before IF:

$$s_i = \frac{1}{M}\sum_{m=1}^{M} -\text{IF}_m\big(\phi_m(\mathbf{r}_{\text{norm},i})\big) \quad (9)$$

Pro: Non-linear anomaly detection via random projections. *Con:* M-fold slower than IF.

CWAE-IF-Latent. Fits IF on latent codes instead of residuals: $s_i = -\text{IF}(\mathbf{z}_i)$. *Pro:* Detects anomalies in the learned representation; complementary to residual scoring. *Con:* Low-dimensional latent space (d_z=32) may lack discriminative power for some datasets.

CWAE-IF-Dual. Ensemble of residual and latent scores with max aggregation:

$$s_i = \max\big(\tilde{s}_i^{\text{res}},\ \tilde{s}_i^{\text{lat}}\big) \quad (10)$$

where $\tilde{s}$ denotes min-max normalized scores from independent IF models. *Pro:* Catches anomalies detectable in either representation. *Con:* Requires fitting two IF models.

CWAE-DIF-Dual. Combines DIF with dual-space aggregation—DIF on both residuals and latent codes, then max aggregation. *Pro:* Most expressive variant (non-linear projections + dual space). *Con:* Heaviest computation ($2M$ projections + $2M$ IF calls).

D. Loss Variants

MSE (default). Mean squared reconstruction error: $\mathcal{L}_{\text{recon}} = \frac{1}{B}\sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$. *Pro:* Simple, smooth gradients everywhere. *Con:* Quadratic penalty amplifies the influence of outliers, which are common in financial data.

CRPS [10]. Continuous Ranked Probability Score under a Gaussian assumption:

$$\mathcal{L}_{\text{CRPS}} = \hat{\sigma}\left[z\big(2\Phi(z)-1\big) + 2\phi(z) - \frac{1}{\sqrt{\pi}}\right] \tag{11}$$

where $z = (\mathbf{x} - \hat{\mathbf{x}})/\hat{\sigma}$, $\hat{\sigma}$ is the batch residual standard deviation, and Φ, ϕ are the standard normal CDF and PDF. *Pro:* Robust to outliers; proper scoring rule with probabilistic interpretation. *Con:* Requires estimating $\hat{\sigma}$ per batch; slightly slower.

Huber [11]. Smooth blend of L_2 near zero and L_1 in the tails:

$$\mathcal{L}_{\text{Huber}}(r,\delta) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le \delta \\ \delta\big(|r| - \frac{\delta}{2}\big) & \text{otherwise} \end{cases} \tag{12}$$

Pro: Smooth gradient at origin (good for optimization) while being robust to large residuals. *Con:* Threshold δ is a hyperparameter that requires tuning.

E. Kernel Variants

Standard IMQ. Bandwidth scales $\{0.2, 0.5, 1.0, 2.0, 5.0\}$. Sufficient for well-behaved latent distributions where most mass is concentrated near the origin.

IMQ-Extreme. Extended scales $\{0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0\}$. *Pro:* Captures similarities at extreme distances, better for heavy-tailed latent spaces. *Con:* More kernel evaluations per batch.

F. Preprocessing Variants

None (default). Standard scaling only.

Tail Compression. A monotonic, invertible transform that compresses values above a quantile threshold τ:

$$\tilde{x} = \begin{cases} x & \text{if } x \le \tau \\ \log(x - \tau + 1) + \tau & \text{if } x > \tau \end{cases} \tag{13}$$

Pro: Reduces reconstruction loss variance on heavy-tailed financial data (e.g., transaction amounts spanning \$0–\$10M). *Con:* May compress informative tail structure if anomalies manifest as extreme values.

G. Baselines

CVAE [3]: Stochastic encoder producing $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ with KL regularization. Requires learning per-sample variance, which can diverge on skewed data. **IF** [7]: Isolation Forest on behavioral features only (ignores context entirely). **IF-concat**: IF on concatenated $[\mathbf{x}; \mathbf{c}]$ (treats context as additional features rather than a conditioning variable). **DIF** [9]: Deep Isolation Forest with frozen random MLP projections on $[\mathbf{x}; \mathbf{c}]$.

III. Experiments

We evaluate 16 methods across 4 synthetic and 11 real datasets, reporting mean AUROC over 3 seeds $\{42, 123, 456\}$.

Synthetic results (Table I). On *Linear* and *Nonlinear* benchmarks, where anomalies are contextual (normal globally, unusual for their context), CWAE-MMD variants achieve 0.988–1.000 AUROC, outperforming IF (0.833) by up to 19 points. CWAE-IF-Dual and CWAE-DIF-Dual reach perfect 1.000 on Nonlinear, demonstrating that dual-space scoring captures complex anomaly structure. On *Scale* (context-dependent variance), all methods struggle; DIF leads at 0.685. On *Multimodal* (categorical context with swap anomalies), IF achieves near-perfect 0.998—these are effectively global anomalies detectable without context modeling, confirming that contextual methods provide no advantage when anomalies are not context-dependent.

Real-world results (Table II). No single method dominates across all 11 datasets. IF-concat wins on 5 datasets (Thyroid, Arrhythmia, Pima, SAML-D, IEEE-CIS), DIF on 2 (Cardio, Ionosphere), and CWAE-MMD variants on 4 (WBC, Vowels, PaySim, CreditCard).

CWAE-CRPS emerges as the strongest CWAE variant, achieving best-in-class on WBC (0.978) and Vowels (0.925). On *PaySim*—a financial fraud dataset with context-dependent transaction patterns—CWAE variants (0.898–0.913) substantially outperform all baselines including IF-concat (0.842) and IF (0.789), providing the strongest evidence for contextual modeling in fraud detection. On CreditCard, several CWAE variants (0.957–0.959) edge out baselines. Conversely, on datasets where anomalies are global rather than contextual (Thyroid, SAML-D), IF-concat suffices and CWAE-MMD provides no advantage.

Notably, plain IF (ignoring context) is consistently the weakest method, ranking last or near-last on 8 of 11 real datasets. This confirms that incorporating context features—whether by concatenation or conditional modeling—is important for real-world anomaly detection.

IV. Discussion and Conclusion

CWAE-MMD provides the largest advantage when anomalies are contextual: 0.99+ AUROC on synthetic benchmarks where IF achieves 0.83. On real datasets with context-behavior structure (PaySim, Vowels, WBC), CWAE-MMD consistently outperforms baselines. When anomalies are global (Thyroid, SAML-D), simpler methods like IF-concat suffice.

Variant guidance. Based on our evaluation: (1) CWAE-CRPS is the recommended default—its robust loss handles heavy-tailed data while maintaining competitive performance across settings; (2) CWAE-DIF-Dual for high-dimensional behavioral features where non-linear residual structure is expected; (3) IF-concat when anomalies are suspected to be global rather than contextual. Across all 15 datasets, CWAE-MMD produced zero NaN training runs, confirming its numerical stability advantage over CVAE.

Limitations. CWAE-MMD requires a training phase, unlike tree-based methods that operate in a single pass. The two-stage pipeline (train autoencoder, then fit IF) introduces additional

TABLE I
AUROC ON SYNTHETIC CONTEXTUAL DATASETS (MEAN OVER 3 SEEDS). **BOLD** = BEST PER COLUMN.

Method	Linear	Scale	Nonlinear	Multimodal
IF	.833	.636	.832	**.998**
IF-concat	.833	.636	.832	**.998**
DIF	.977	**.685**	.987	.838
CVAE	.953	.673	.989	.591
CWAE-IF	.988	.634	.997	.641
CWAE-DIF	.988	.632	.997	.656
CWAE-IF-Latent	.994	.675	.996	.587
CWAE-IF-Dual	**.995**	.662	**1.000**	.585
CWAE-DIF-Dual	.994	.664	**1.000**	.594
CWAE-CRPS	.990	.644	.999	.661
CWAE-CRPS-IF	.990	.644	.999	.661
CWAE-Huber	.985	.664	.999	.731
CWAE-IMQ-Ext	.981	.649	.998	.652
CWAE-Beta0.1	.981	.636	.995	.723
CWAE-Beta10	.980	.652	.993	.612
CWAE-Latent16	.985	.659	.996	.613

TABLE II
AUROC ON REAL-WORLD DATASETS (MEAN OVER 3 SEEDS). **BOLD** = BEST PER COLUMN.

Method	*Thyroid*	*Cardio*	*Arrhyth.*	*Ionosph.*	*Pima*	*WBC*	*Vowels*	*SAML-D*	*PaySim*	*CreditCd*	*IEEE-CIS*
IF	.962	.855	.784	.852	.651	.966	.662	.514	.789	.949	.606
IF-concat	**.990**	.954	**.794**	.908	**.723**	.958	.753	**.703**	.842	.949	**.642**
DIF	.964	**.964**	.759	**.940**	.708	.959	.843	.683	.778	.947	.586
CVAE	.879	.725	.779	.919	.662	.964	.838	.618	.873	.952	.561
CWAE-IF	.890	.842	.787	.934	.672	.974	.906	.666	**.913**	.957	.592
CWAE-CRPS	.928	.816	.782	.931	.698	**.978**	**.925**	.629	.898	.956	.639
CWAE-Huber	.886	.814	.786	.934	.664	.970	.899	.640	.911	.959	.613
CWAE-IF-Lat	.964	.872	.723	.616	.617	.911	.702	.639	.778	.952	.552
CWAE-IF-Dual	.956	.879	.787	.823	.675	.966	.861	.672	.902	.959	.593
CWAE-DIF-D	.954	.885	.771	.882	.680	.959	.882	.672	.910	**.959**	.603

hyperparameters. On datasets where context does not explain behavioral variation (e.g., SAML-D), the conditional model provides no benefit over simpler approaches.

An extended version with AUPRC analysis, scalability studies, and ablation experiments is in preparation.

REFERENCES

[1] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," *Conference on Credit Scoring and Credit Control*, 2001.

[2] Financial Action Task Force, "International standards on combating money laundering," FATF/OECD, Tech. Rep., 2012.

[3] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, vol. 28, 2015.

[4] A. A. Pol, V. Berber, C. Germain, G. Cerminara *et al.*, "Anomaly detection with conditional variational autoencoders," in *2019 18th IEEE ICMLA*, 2019, pp. 1651–1657.

[5] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," in *International Conference on Learning Representations*, 2018.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schoelkopf, and A. J. Smola, "A kernel two-sample test," *JMLR*, vol. 13, pp. 723–773, 2012.

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2014.

[9] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep isolation forest for anomaly detection," *IEEE TKDE*, vol. 35, no. 12, pp. 12 591–12 604, 2023.

[10] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *JASA*, vol. 102, no. 477, pp. 359–378, 2007.

[11] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.