Causal Prototype Attention Classifier: An Interpretable Model for Credit Card Fraud Detection Under Extreme Class Imbalance

Claudio Giusti[1], Mirko Casu[1], Luca Guarnera[1] and Sebastiano Battiato[1]

[1]*Department of Mathematics and Computer Science, University of Catania, Italy*

Abstract

Detecting fraud in real-world financial systems presents unique challenges due to the extreme imbalance between legitimate and fraudulent transactions. This imbalance not only hampers the performance of standard classifiers but also calls for models that are both effective and interpretable. In this work, we propose the Causal Prototype Attention Classifier (CPAC), an interpretable model tailored for binary classification under severe class imbalance. CPAC learns two class prototypes in latent space and uses a per-feature attention mechanism to compute weighted distances from each input to the prototypes. This architecture enables interpretable, prototype-based reasoning instead of relying on opaque decision boundaries. We evaluate CPAC on a credit card fraud detection task, comparing it to standard classifiers, Logistic Regression, Random Forest, and XGBoost, under two widely adopted oversampling strategies: SMOTE and a VAE–GAN-based generator. Results show that CPAC achieves performance comparable to black-box models while offering greater transparency and stability, particularly when trained on VAE–GAN augmented data. These findings support the adoption of CPAC in applications where both performance and interpretability are essential.

Keywords

Fraud Detection, Imbalanced Learning, Prototype-Based Models, Interpretable Machine Learning, VAE-GAN, SMOTE

1. Introduction

Fraud detection in financial systems faces significant challenges due to extreme class imbalance between legitimate and fraudulent transactions. This imbalance, common in forensic applications like deepfake detection [1, 2, 3], causes systematic misclassifications and cognitive biases such as the 'impostor bias' [4]. Traditional models including Logistic Regression [5], Random Forests [6], and XGBoost [7] perform well on tabular data but struggle with severe imbalance, requiring oversampling or cost-sensitive methods. Two oversampling approaches dominate: SMOTE-based interpolation [8], which may generate unrealistic samples, and generative models like VAEs [9] and GANs [10] that capture richer distributions. However, these traditional methods lack interpretability, limiting their use in high-stakes domains. Our study centers on the Causal Prototype Attention Classifier (CPAC), an interpretable architecture designed to address these challenges. Our contributions:

- We propose CPAC, a prototype-based model with feature-wise attention for extreme imbalance.
- We benchmark CPAC and standard classifiers (Logistic Regression, Random Forest, XGBoost) on the Credit Card Fraud Detection dataset under both SMOTE and VAE–GAN oversampling.
- We show that CPAC matches or surpasses black-box models on realistic synthetic data, maintaining interpretability and scalability.

Section 2 presents our methodological pipeline, while Section 3 reports experiments and comparative results.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

claudio.giusti@studium.unict.it (C. Giusti); mirko.casu@phd.unict.it (M. Casu); luca.guarnera@unict.it (L. Guarnera); sebastiano.battiato@unict.it (S. Battiato)

https://github.com/claudiunderthehood (C. Giusti)

0000-0001-6975-2241 (M. Casu); 0000-0001-8315-351X (L. Guarnera); 0000-0001-6127-2470 (S. Battiato)

Related Work

Extreme class imbalance in fraud detection has been traditionally tackled with data-level solutions. SMOTE [11]. More recently, deep generative approaches such as VAE and GAN hybrids, have been used to synthesize realistic anomalies, improving rare fraud detection [12]. Interpretability has also advanced: prototype-based networks like ProtoPNet [13]. Attention mechanisms are widely used for feature weighting, though their value as explanations is debated [14, 15]. Finally, model-agnostic methods like LIME [16]. provide local interpretability via surrogate models and feature attributions.

2. Methodologies

This work will focus mainly on introducing the Causal Prototype Attention Classifier (CPAC) as a new way to detect frauds and compare its performances to state-of-the-art classifiers. We will test the CPAC against Logistic Regression, Random Forest Classifier and XGBoost; all four of them will be tested with the firstly with the non-augmented dataset, in order to establish whether there were any improvements, and then they will be tested with augmented dataset by SMOTE and VAE-GAN to compare the performances and determine whether the CPAC offers a good alternative against standard models.

Dataset and Preprocessing

All experiments use the public Credit Card Fraud Detection dataset from Kaggle, released by Worldline and ULB, containing 284,807 anonymized European cardholder transactions from September 2013. Only 492 entries are fraudulent (0.17%), resulting in an extreme class imbalance. Each transaction is described by 30 features: 28 principal components (PCA-transformed for privacy), plus Time and Amount. The target `Class` is binary (1 for fraud, 0 otherwise). To ensure fair model training, features are robustly normalized using the median and interquartile range (IQR):

$$x_{\text{norm}} = \frac{x - \tilde{x}}{\text{IQR}(x)}$$

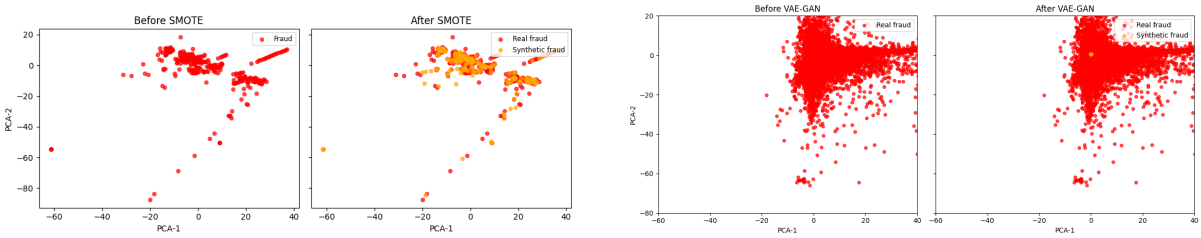
where the median $\tilde{x}$ and IQR are computed for each feature. If $\text{IQR}(x) = 0$, we use a default unit divisor for stability. Finally, we apply stratified sampling to split the dataset into 70% training and 30% validation sets, preserving class balance across splits.

Oversampling Strategies

To address class imbalance, we compared two oversampling methods: SMOTE and a custom VAE-GAN pipeline. SMOTE interpolates between minority samples to synthetically expand the fraud class, reaching 300, 600, and 900 total fraud cases; these are merged, shuffled, and split for training and validation (Figure 1a). In parallel, our VAE-GAN, trained solely on frauds, generates new samples by decoding random latent vectors, resulting in more clustered, nonlinear fraud distributions (Figure 1b). While SMOTE covers existing clusters uniformly, sometimes producing unrealistic samples in sparse regions, VAE-GAN requires more tuning but yields diverse, manifold-aware frauds. We trained all classifiers and CPAC on both types of augmented data to assess their impact.

Baseline Classifiers

We benchmarked CPAC against three standard models: Logistic Regression, Random Forest, and XGBoost. **Logistic Regression**, a linear classifier, performed well with SMOTE-augmented data but struggled to separate the more complex distributions produced by VAE-GAN. **Random Forest**, an ensemble of decision trees, achieved high precision, though it showed signs of overfitting with large synthetic datasets, especially at higher sample counts. **XGBoost** consistently delivered the best and



(a) PCA plots comparing frauds distribution before and after SMOTE oversampling.

(b) PCA plots comparing frauds distribution before and after VAE-GAN oversampling.

Figure 1: Comparison of fraud distribution using different oversampling techniques.

most stable metrics across all settings, though its improvements plateaued with extensive oversampling. All classifiers were tested on both original and augmented datasets to assess generalization.

Causal Prototype Attention Classifier (CPAC)

The Causal Prototype Attention Classifier (CPAC) introduces interpretable, class-aware reasoning into neural classification. It learns two prototype vectors $\mathbf{p}_0, \mathbf{p}_1 \in \mathbb{R}^d$ representing the centroids of the normal and fraud classes. An attention module assigns per-feature importance weights $\mathbf{w} \in (0, 1)^d$ to highlight relevant dimensions. The term "causal" here is used loosely to suggest that the attention weights may help highlight which latent features have a higher impact on the outcome. Classification is performed by computing the weighted distances of an input $\mathbf{x}$ to each prototype, and interpreting the negative distances as logits for a softmax-based prediction. Its structure can be visualised in Figure 2.

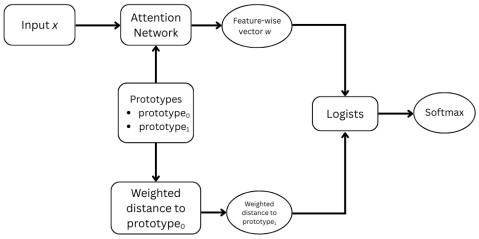


Figure 2: Architecture of the CPAC model. Each input is compared to class prototypes using an attention-weighted distance, followed by softmax scoring.

Training Loss

To handle class imbalance and focus on hard fraud cases, CPAC is trained using the Focal Loss:

$$\mathcal{L}_{\text{FL}}(y, \hat{y}) = -\,\alpha\,(1-\hat{y})^{\gamma}\,y\,\log\hat{y} \quad -(1-\alpha)\,\hat{y}^{\gamma}\,(1-y)\,\log(1-\hat{y}). \tag{1}$$

We use $\alpha = 0.95$ to favor the minority (fraud) class and $\gamma = 2.0$ to down-weight well-classified examples. This helps CPAC focus learning on the most ambiguous, fraud-relevant patterns. In order to avoid wasting training time early-stopping was introduced; it uses a composite score that slightly prioritizes Precision over Recall.

3. Experiments

In this section we will finally test the CPAC metrics against the Logistic Regression, Random Forest and XGBoost and evaluate what it does best, what it does worse and overall if it is worth to be considered as a reliable method to detect frauds overall. Both SMOTE and the VAE-GAN generate synthetic frauds that get concatenated to the original dataset.

No Oversampling

We tested the models on the non augmented dataset. Without augmentation, as we can see in Table 1, the Logistic Regression struggles the most to identify which is fraudulent and which is not, meaning that oversampling could aid its training and boost its performances. Random Forest achieves the highest precision with a strong recall metric, suggesting that its performances could only benefit with oversampling. XGBoost seems the most stable, achieves the highest recall (not considering Logistic Regression because of its instability, like we will do with VAE-GAN oversampling) and as well as the Random Forest, its performances will only improve with an enriched dataset. The CPAC, despite being more stable than the Logistic Regression, still struggles to reach metrics comparable to the previous two, clearly urging for a richer dataset to help its prototypes find the right anchor point in its representation.

Table 1
Benchmark results on the original (non-augmented) dataset.

Model	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC (%)
Logistic Regression	56.67	92.76	61.32	96.79
Random Forest	**98.87**	80.07	87.22	92.75
XGBoost	96.70	**88.51**	**92.21**	**96.80**
CPAC	87.20	73.65	79.85	95.80

SMOTE Oversampling Performances

As shown in Table 2, Logistic Regression, Random Forest, and XGBoost already perform strongly with 300 synthetic samples, achieving high scores across all metrics, while CPAC initially lags behind. At 600 samples, CPAC sees a significant gain in precision, although its recall remains lower than the other three. SMOTE does not align well with CPAC's architecture, which benefits from a more clustered latent space as produced by generative models like VAE-GAN. By 900 samples, CPAC shows further improvement, while the other classifiers begin to plateau. This suggests that, unlike the other models which may saturate or overfit with additional data, CPAC could continue to benefit from further oversampling.

Table 2
Benchmark results using SMOTE oversampling with 300, 600, and 900 synthetic fraud samples.

# Samples	Model	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC (%)
300	Logistic Regression	95.01	86.49	90.30	98.40
300	Random Forest	**98.52**	92.19	95.13	98.20
300	XGBoost	98.30	**92.61**	**95.28**	97.99
300	CPAC	89.91	86.50	88.17	**99.30**
600	Logistic Regression	94.98	90.04	92.37	98.08
600	Random Forest	**98.77**	93.27	95.85	97.73
600	XGBoost	98.14	**94.03**	**95.99**	**98.60**
600	CPAC	94.43	88.07	91.14	97.40
900	Logistic Regression	95.68	88.59	91.84	98.39
900	Random Forest	**98.44**	92.08	95.04	98.18
900	XGBoost	97.94	**92.68**	**95.16**	**99.27**
900	CPAC	94.36	92.33	93.33	98.90

VAE-GAN Oversampling Performances

In Table 3, we evaluate the models using VAE-GAN as an oversampler. In this setting, CPAC begins with strong performance right from 300 samples, outperforming Random Forest in recall and closely trailing XGBoost. This is due to the architectural synergy between CPAC and the VAE-GAN, which generates clustered, non-linear samples that align well with CPAC's prototype-based structure. In contrast, Logistic Regression performs poorly; as a linear classifier, it struggles to separate complex manifolds and clustered distributions, explaining its consistently lower scores across all sample sizes. At 600 samples, CPAC achieves the highest recall and AUC among all models, indicating it is beginning to internalize a more structured latent representation of the classes. By 900 samples, both Random Forest and XGBoost show signs of saturation, with marginal or no improvements, while CPAC continues to

gain, ultimately outperforming all models in both recall and AUC. These results reinforce the trend observed in the SMOTE experiments: while traditional classifiers may plateau or overfit with increasing synthetic data, CPAC benefits from larger, generatively crafted datasets by continuing to refine its internal structure.

Table 3
Benchmark results using a VAE–GAN oversampler with 300, 600, and 900 synthetic fraud samples.

# Samples	Model	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC (%)
300	Logistic Regression	73.98	95.03	81.27	95.03
300	Random Forest	**98.92**	89.07	93.44	89.07
300	XGBoost	98.06	**92.22**	**94.95**	92.22
300	CPAC	94.12	91.16	92.59	**97.79**
600	Logistic Regression	68.86	96.65	76.77	96.65
600	Random Forest	**99.27**	92.68	95.74	92.68
600	XGBoost	99.15	95.27	**97.13**	95.27
600	CPAC	95.84	**95.56**	95.70	**97.67**
900	Logistic Regression	74.93	97.49	82.64	97.49
900	Random Forest	**99.71**	95.09	97.29	95.09
900	XGBoost	98.97	96.29	**97.59**	96.29
900	CPAC	96.63	**96.51**	96.57	**98.27**

Final Observations

Across both oversampling techniques, CPAC consistently improves as more high-quality synthetic frauds are added, particularly when they are generated via VAE-GAN. While standard classifiers demonstrate strong early performance, they tend to saturate quickly. CPAC, on the other hand, shows a more stable and scalable learning curve, especially in terms of recall and AUC. These findings suggest that CPAC is not only competitive but also more resilient to overfitting in high-imbalance scenarios, making it a promising architecture for fraud detection in data-scarce or synthetic-rich environments.

4. Conclusions

In this work, we presented CPAC as a promising and competitive approach to fraud detection compared to standard classifiers. We first introduced its architecture and training process, then benchmarked its performance against three standard models (Logistic Regression, Random Forest, and XGBoost) on datasets augmented using SMOTE and VAE-GAN. Ultimately, we demonstrated that CPAC not only improves with more samples but also has the potential to surpass other models with further training, highlighting its stability and resilience.

Acknowledgments

This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Z. Ke, S. Zhou, Y. Zhou, C. H. Chang, R. Zhang, Detection of AI deepfake and fraud in online payments using gan-based models, arXiv preprint arXiv:2501.07033 (2025). URL: https://arxiv.org/abs/2501.07033.

[2] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, et al., Deepfake media forensics: Status and future challenges, Journal of Imaging 11 (2025) 73.

[3] A. Sharma, R. Tiwari, Banking in the Age of Deepfakes: Evaluating Perceptions of Deepfake Fraud Risks, in: Navigating the World of Deepfake Technology, IGI Global Scientific Publishing, 2024, pp. 454–469. doi:10.4018/979-8-3693-5298-4.ch023.

[4] M. Casu, L. Guarnera, P. Caponnetto, S. Battiato, GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions, Forensic Science International: Digital Investigation 50 (2024) 301795. doi:https://doi.org/10.1016/j.fsidi.2024.301795.

[5] D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Wiley, 2013. doi:10.1002/9781118548387.

[6] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[7] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system (2016) 785–794. doi:10.1145/2939672.2939785.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357. doi:10.1613/jair.953.

[9] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, arXiv preprint arXiv:1312.6114 (2014). URL: https://arxiv.org/abs/1312.6114.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953.

[12] T. Tang, J. Yao, Y. Wang, Q. Sha, H. Feng, Z. Xu, Application of Deep Generative Models for Anomaly Detection in Complex Financial Transactions, arXiv preprint arXiv:2504.15491 (2025). URL: https://arxiv.org/abs/2504.15491.

[13] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su, This Looks Like That: Deep Learning for Interpretable Image Recognition, in: Advances in Neural Information Processing Systems, editor = H. Wallach and H. Larochelle and A. Beygelzimer and F. d'Alché-Buc and E. Fox and R. Garnett, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

[14] S. Jain, B. C. Wallace, Attention is not Explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. doi:10.18653/v1/N19-1357.

[15] S. Wiegreffe, Y. Pinter, Attention is not not Explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20. doi:10.18653/v1/D19-1002.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. doi:10.1145/2939672.2939778.