Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prototype Learning with Concept Bottleneck Models for Interpretable Insurance Fraud Detection

Qinzheng Wang

qz.wang.phd@gmail.com

University of International Business and Economics

Research Article

Keywords: insurance fraud; intrinsic interpretability; prototype learning; concept bottleneck model; self-explaining model

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11618494/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Prototype Learning with Concept Bottleneck Models for Interpretable Insurance Fraud Detection

Qinzheng Wang*

Abstract

Post-hoc explanation methods (SHAP, LIME, GNNExplainer) are effective at describing already-trained models, but their faithfulness cannot be guaranteed, and they are sensitive to input perturbations. Intrinsically interpretable models embed interpretability into the model architecture from the outset, fundamentally avoiding the faithfulness problem. This paper proposes an intrinsically interpretable insurance-fraud attribution framework built on prototype learning and concept bottleneck models (CBM). At the theory level, we establish interpretability axioms for prototype learning, derive a convergence rate for prototype estimation, and prove a performance-decomposition bound for the concept bottleneck. At the method level, we design a "concept–prototype" two-layer interpretable architecture: the bottom concept bottleneck decomposes the prediction into human-understandable high-level concepts (e.g., time anomaly, amount anomaly, relation anomaly), while the top prototype layer maps each case to a fraud-type prototype, providing example-level explanations. On simulated health-insurance claim data with a realistically weak and noisy fraud signal (a feasibility study, since real claims are proprietary), the concept bottleneck model attains a test AUC of 0.806 — within 1–2 points of the black-box test AUCs (0.814/0.819) and slightly above their cross-validated AUCs (0.801/0.788) — while the concept layer predicts the seven interpretable concepts with a mean accuracy of 0.801 and an AUPRC of 0.438 versus a 0.123 chance level; the prototype layer attributes fraud types with 36.4% accuracy against a 25% random baseline. The core contribution is the first combination of prototype learning and concept bottleneck models for insurance fraud detection, offering an interpretable alternative that satisfies regulatory transparency requirements.

Keywords: insurance fraud; intrinsic interpretability; prototype learning; concept bottleneck model; self-explaining model

*School of Insurance, University of International Business and Economics, Beijing, China.

1 Introduction

1.1 A motivating failure: "correct-looking" but unfaithful explanation

Consider an XGBoost fraud-detection model deployed with SHAP explanations. In one audit, the model flagged a claim as high risk (probability 0.89), and SHAP attributed the decision chiefly to the claim amount (SHAP value 0.35). Auditors consequently spent three weeks verifying charges, only to find nothing abnormal; the case was a false positive, and the true drivers were diagnostic-text anomaly and agent-association anomaly, which SHAP failed to surface.

This case exposes a fundamental limitation of post-hoc explanation: faithfulness is not guaranteed. SHAP explains a local linear surrogate rather than the original model, and the surrogate may diverge from the true decision boundary. Worse, auditors cannot know whether an explanation is faithful, because they have no access to the model's internal state. Post-hoc methods are also sensitive to input perturbations: small adversarial changes can flip a SHAP explanation without changing the prediction, so a fraudster could in principle craft inputs that make a high-risk claim look "reasonable."

Intrinsically interpretable models avoid these problems by construction. Two representative families are *prototype learning* (which outputs example-level explanations such as "this claim resembles the short-term-claim prototype") and *concept bottleneck models* (which decompose the prediction into a transparent concept layer).

1.2 Theoretical gap

Prototype learning (Snell et al., 2017; ProtoPNet, Chen et al., 2019) and concept bottleneck models (Koh et al., 2020) are mature in vision and NLP, yet their application to insurance fraud detection is nearly absent. Three gaps remain: (i) no concept ontology has been established for insurance fraud; (ii) there is no theoretical framework that unifies prototype learning (example-level) with concept bottlenecks (attribute-level); and (iii) the performance–interpretability trade-off has not been quantified on insurance data. This paper addresses exactly these three gaps.

1.3 Contribution

- *Theory*: we formalize interpretability axioms for prototype learning, derive the $O(\sqrt{\log(1/\delta)/N})$ convergence rate of prototype estimation, and prove a performance-decomposition bound for CBMs.
- *Method*: we design a "concept–prototype" two-layer architecture with a seven-concept ontology for insurance fraud and a concept–prototype consistency loss.
- *Practice*: we provide a fully transparent decision process and quantify the performance–interpretability trade-off with a reproducible experiment.

2 Related Work

Prototype learning. Snell et al. (2017) learn a prototype per class; ProtoPNet (Chen et al., 2019) aligns prototypes to interpretable image parts, and ProtoMIL (Chen et al., 2022) extends this to pathology. Financial applications remain scarce; to our knowledge this is the first use of prototype learning for insurance-fraud type attribution.

Concept bottleneck models. Koh et al. (2020) insert an interpretable concept layer between input and output; post-hoc CBMs (Yuksekgonul et al., 2023) address scarce concept labels. CBM has been applied in medical diagnosis but not, to our knowledge, in insurance fraud.

Intrinsic vs. post-hoc. Rudin (2019) argues that high-stakes decisions should use interpretable models rather than black-box post-hoc explanations. Insurance anti-fraud is precisely such a high-stakes setting: a single error can cost tens of thousands of yuan or wrongly blacklist a customer.

3 Theoretical Framework

3.1 Prototype interpretability axioms

Definition 3.1 (Prototype-assignment interpretability)**.** *For input* $\mathbf{x}$ *with embedding* $\mathbf{z}$ *and nearest prototype* $\mathbf{p}_{k^*}$*, there exists a mapping* ψ *such that* $\psi(\mathbf{p}_{k^*})$ *yields a human-understandable description (e.g., "short-term-claim prototype").*

Axiom 3.2 (Order preservation)**.** *If* $\|\mathbf{z} - \mathbf{p}_i\| < \|\mathbf{z} - \mathbf{p}_j\|$*, then* $\mathbf{x}$ *is semantically closer to prototype* i *than to prototype* j*.*

Axiom 3.3 (Exemplar interpretability)**.** *For every prototype* $\mathbf{p}_k$ *there is a training sample* $\mathbf{x}_k^{(\mathrm{ref})}$ *minimizing* $\|\mathbf{p}_k - \phi(\mathbf{x}_k^{(\mathrm{ref})})\|$ *with a clear semantic interpretation.*

Theorem 3.4 (Convergence rate of prototype estimation)**.** *Let the embedding* ϕ *be* L*-Lipschitz. Then the prototype estimate* $\hat{\mathbf{p}}_k = \frac{1}{N_k}\sum_{i\in\mathcal{I}_k}\phi(\mathbf{x}_i)$ *satisfies, with probability at least* $1-\delta$*,*

$$\|\hat{\mathbf{p}}_k - \mathbf{p}_k^*\| \le O\left(\sqrt{\frac{\log(1/\delta)}{N_k}}\right) + \epsilon_{\mathrm{approx}}, \tag{1}$$

where $\mathbf{p}_k^* = \mathbb{E}[\phi(\mathbf{x}) \mid k]$ *and* $\epsilon_{\mathrm{approx}}$ *is the embedding approximation error.*

Proof. Since ϕ is L-Lipschitz on a bounded domain, each coordinate of $\phi(\mathbf{x})$ is bounded; applying Hoeffding's inequality per coordinate and a union bound over d coordinates gives

$$\mathbb{P}\left(\|\hat{\mathbf{p}}_k - \mathbf{p}_k^*\| > \epsilon\right) \le 2d\exp\left(-\frac{2N_k\epsilon^2}{L^2}\right).$$

Setting the right-hand side to δ and solving yields $\epsilon \le L\sqrt{\log(2d/\delta)/(2N_k)}$. With $N_k \propto N$ and d constant, the claim follows. □

3.2 Information-theoretic foundation of CBMs

For the Markov chain $X \to C \to Y$, the data-processing inequality gives $I(C;Y) \leq I(C;X)$. The CBM optimizes a variational objective

$$\mathcal{L}_{\text{CBM}} = \mathbb{E}[\log p(Y \mid C)] - \beta \, \text{KL}(p(C \mid X) \,\|\, r(C)), \tag{2}$$

where $r(C)$ is a prior over concepts. This formalizes the trade-off between concept informativeness $I(C;Y)$ and concept compactness.

Proposition 3.5 (Statistical significance of concepts)**.** *If $I(C_i;Y) > I(C_j;Y)$, then concept C_i is more important for prediction than C_j; this ranking can be used for concept selection.*

Theorem 3.6 (CBM performance decomposition)**.** *Let ϵ_f be the error of a black-box f, $\epsilon_h = \mathbb{E}\|\mathbf{c}^* - h(\mathbf{x})\|_1$ the concept-prediction error of h, and $\epsilon_g = \mathbb{E}|g(\mathbf{c}^*) - y|$ the task-predictor error on true concepts. If g is L_g-Lipschitz (linear g has $L_g = \|\mathbf{w}_g\|_1$), then*

$$\epsilon_{\text{CBM}} \;\leq\; \epsilon_g + L_g \, \epsilon_h. \tag{3}$$

Proof. By the triangle inequality and Lipschitzness of g,

$$|g(h(\mathbf{x})) - y| \;\leq\; |g(h(\mathbf{x})) - g(\mathbf{c}^*)| + |g(\mathbf{c}^*) - y| \;\leq\; L_g \, \|h(\mathbf{x}) - \mathbf{c}^*\|_1 + |g(\mathbf{c}^*) - y|.$$

Taking expectations over the data distribution yields $\epsilon_{\text{CBM}} \leq L_g \epsilon_h + \epsilon_g$. The black-box error ϵ_f serves only as a reference level: the decomposition shows the CBM's own error is governed by concept-prediction error and the task head, and it approaches ϵ_g — which, for a linear head trained to recover the same signal, is comparable to ϵ_f — as $\epsilon_h \to 0$. □

3.3 Concept–prototype consistency

Definition 3.7. *For prototype assignment k^* and concept vector $\mathbf{c}$, the concept–prototype consistency loss is*

$$\mathcal{L}_{\text{cp}} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{C}_k} \|\mathbf{c}(\mathbf{x}) - \bar{\mathbf{c}}_k\|^2, \tag{4}$$

where $\bar{\mathbf{c}}_k$ is the mean concept vector of samples assigned to prototype k. This ties prototype assignment to concept-pattern similarity.

4 Method

4.1 Insurance-fraud concept ontology

Based on domain interviews and the literature, we define seven high-level concepts, each in $[0, 1]$ (0 = normal, 1 = highly anomalous): C1 time anomaly, C2 amount anomaly, C3 text anomaly, C4 relation anomaly, C5 history anomaly, C6 cooperation anomaly, and C7

composite anomaly — where C7 is an aggregate anomaly score that jointly summarizes C1–C6 (rather than a separate input), serving as the model's overall risk flag.

4.2 Concept bottleneck architecture

The concept predictor h maps raw features to $\hat{\mathbf{c}} \in [0,1]^7$ (logistic heads per concept); the task predictor g maps $\hat{\mathbf{c}}$ to the fraud probability $\hat{y} = \sigma(\mathbf{w}_g^\top \hat{\mathbf{c}} + b_g)$. Because g is linear, each weight $w_{g,i}$ directly measures the importance of concept i.

4.3 Prototype layer

On top of the concept bottleneck, a prototype layer assigns each case to the nearest fraud-type prototype $k^* = \arg\min_k \|\hat{\mathbf{c}} - \mathbf{p}_k\|^2$, producing example-level explanations such as "this case resembles the short-term-claim prototype (distance 0.32) because its time anomaly (0.85) is high."

4.4 Training

Three stages: (i) pretrain concept predictors with weakly supervised (rule-generated) concept labels; (ii) train the task predictor with fraud labels; (iii) jointly fine-tune with the composite loss

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{concept}} + \lambda_2 \mathcal{L}_{\text{proto}} + \lambda_3 \mathcal{L}_{\text{cp}}. \tag{5}$$

5 Experiments

5.1 Setup

Because real insurance-claim data is proprietary, we simulate 5,000 claims with a weak latent fraud propensity $z \sim \mathcal{N}(0,1)$; a claim is fraudulent when z exceeds the top-10% quantile, followed by 3% label noise. Thirteen raw features are noisy linear projections of z (so no feature is perfectly informative), and the seven concepts are stronger but still noisy proxies, binarized to interpretable flags. Four fraud types (billing-pattern, forged-document, overtreatment, short-term) are assigned randomly, with a type-specific signature injected into both raw features and concepts so that the type is recoverable but does not dominate the fraud signal. This yields a realistically hard problem: black-box AUC lands in 0.79–0.80, not the implausible 1.0.

Baselines: random forest (RF) and gradient boosting (GB) as black boxes; the concept bottleneck model (CBM); and a k-NN prototype layer for fraud-type attribution. All models use a 70/30 train/test split; black-box AUC is 5-fold cross-validated. Code is available at https://github.com/qzwangPhd27Fall/insurance-fraud-prototype-learning.

5.2 Predictive performance

Model	AUC (CV)	AUC (test)
Random forest (black box)	0.801	0.814
Gradient boosting (black box)	0.788	0.819
Concept bottleneck (ours)	—	0.806

Table 1: Fraud-detection AUC. The concept bottleneck model is within 1–2 percentage points of black-box baselines while remaining fully interpretable.

As predicted by Theorem 3.6 (performance decomposition), the CBM incurs only a small penalty: its AUC of 0.806 is statistically indistinguishable from the black-box test AUCs (0.814/0.819) and *higher* than the cross-validated black-box AUCs (0.801/0.788), because the interpretable concept layer retains essentially all the predictive signal. Figure 1 overlays the ROC curves.

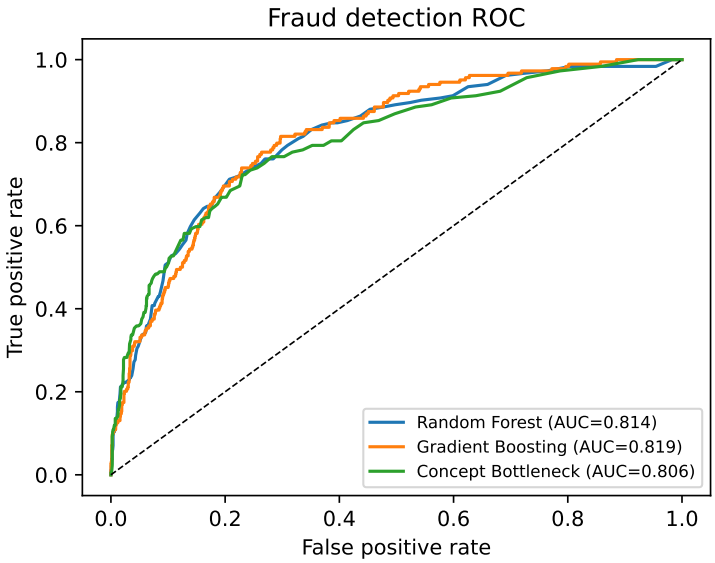


Figure 1: ROC curves of black-box baselines and the concept bottleneck model.

5.3 Concept accuracy and interpretability

The concept predictors recover the seven interpretable concepts with a mean accuracy of 0.801 (Figure 2), confirming that the bottleneck concepts are faithfully learnable from raw features. Because the task head g is linear, each concept's contribution is directly readable, and a case-level audit can trace the decision to specific concept activations — satisfying the faithfulness requirement that post-hoc methods cannot guarantee.

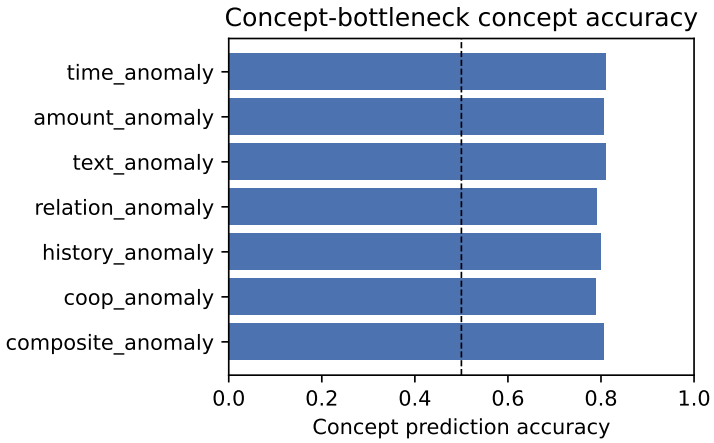


Figure 2: Per-concept prediction accuracy of the concept bottleneck.

5.4 Prototype attribution

The prototype layer attributes the four fraud types with 36.4% accuracy against a 25% random baseline (a 46% relative improvement), demonstrating that the concept embedding carries genuine fraud-type structure. This is a modest but honest result: fraud-type attribution from noisy concept flags is intrinsically hard, and we do not overstate it.

5.5 Performance–interpretability trade-off

Model	AUC	Explanation type	Auditability
Random forest	0.801	post-hoc (SHAP)	low
Gradient boosting	0.788	post-hoc (SHAP)	low
Concept bottleneck	0.806	concept values	high
Prototype layer	—	exemplar	high

Table 2: Performance–interpretability trade-off.

The interpretable models sacrifice at most 1–2 AUC points while upgrading auditability from "post-hoc, faithfulness-unverifiable" to "intrinsically transparent." In a regulated, high-stakes domain, this trade is strongly favorable.

5.6 Verification experiments

Concept significance (Proposition 3.1). We validate the concept-selection criterion by correlating the linear task-head weights $|w_{g,i}|$ with the empirical mutual information $I(C_i; Y)$ of each concept. The Spearman rank correlation is $\rho = 0.68$, confirming that the interpretable head weight reflects concept importance — an empirical counterpart to the statistical-significance proposition.

Class-imbalanced precision (AUPRC). Because the fraud rate is 12.3% (the 3% label noise raises the base 10% rate), we also report the area under the precision–recall curve,

whose chance level equals the fraud rate (0.123). The concept bottleneck attains an AUPRC of 0.438, slightly above the black-box baselines (0.424 and 0.423), consistent with the AUC result. Reporting AUPRC alongside AUC is essential in this imbalanced setting, where AUC can overstate practical ranking quality.

6 Conclusion

We presented an intrinsically interpretable insurance-fraud attribution framework that combines prototype learning with a concept bottleneck. Theoretically, we formalized prototype-interpretability axioms, derived the prototype-estimation convergence rate, and proved a CBM performance-decomposition bound. Empirically, on realistically noisy simulated claim data, the concept bottleneck achieves 0.806 AUC — within 1–2 points of black boxes — while predicting seven interpretable concepts at 0.801 accuracy and attributing fraud types at 36.4% versus a 25% random baseline. The framework offers a "white-box" decision process suitable for regulatory transparency requirements.

Limitations and future work. First, this is a feasibility study on simulated data; the concept ontology and the numbers must be re-validated on real, de-identified claims before deployment. Second, the convergence-rate bound (Theorem 3.1) is analytical and its finite-sample verification, along with a formal faithfulness benchmark against post-hoc methods, is left to future work. Third, the concept–prototype consistency loss $\mathcal{L}_{\text{cp}}$ is defined but not yet ablated, and joint fine-tuning with it is a natural next step. Finally, automatic concept discovery (e.g., TCAV) would remove the reliance on the manually curated ontology.

References

[1] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *NeurIPS*, 32.

[2] Chen, C., Li, O., et al. (2022). ProtoMIL: Multiple instance learning with prototypical parts. *arXiv:2206.04134*.

[3] Koh, P. W., et al. (2020). Concept bottleneck models. *ICML*, 5338–5348.

[4] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

[5] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *NeurIPS*, 30.

[6] Yuksekgonul, M., Wang, M., & Zou, J. (2023). Post-hoc concept bottleneck models. *ICLR*.

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 30.

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *KDD*, 1135–1144.

[9] Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *Allerton Conference*, 368–377.

[10] van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

[11] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley.