Practical Framework for Mineral Prospectivity Mapping: Integrating Uncertainty Quantification with Spatial Block Cross-Validation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Mineral prospectivity mapping (MPM) guides exploration targeting by predicting mineralization potential, yet most machine learning approaches provide only point probability estimates without uncertainty measures or statistical guarantees essential for risk-informed decisions. Current methods face label ambiguity from undiscovered deposits, spatial autocorrelation invalidating standard cross-validation, and absence of practical guidance linking probabilistic outputs to exploration decisions. We present a practical framework that combines positive-unlabeled (PU) learning, nested spatial block cross-validation, and a practical zone classification system. The framework combines bootstrap ensemble uncertainty with cross-conformal prediction to assign each spatial unit to one of five exploration priority tiers, from "immediate" drilling targets to "excluded" regions, using lift-based percentile thresholds and a 30% relative interquartile range (rel_IQR) uncertainty criterion. Validation on the Yilgarn Craton, Western Australia (70,200 grid cells, 818 nickel deposits, 1.17% positive rate) employs nested five-fold spatial block CV with 50 km blocks informed by variogram analysis. Both XGBoost and BaggingPU-XGBoost achieve outer test PR-AUC of 0.195 and ROC-AUC of 0.911. The practical zone classification captures 55% of deposits within 5% of study area (11-fold concentration), while conformal exclusion removes less than 4% of deposits, maintaining 96% empirical coverage. BaggingPU-XGBoost produces more stable high-probability predictions (median rel_IQR 0.18 versus 0.48 for XGBoost in top 1%), enabling confident identification of immediate targets. Our framework bridges the gap between academic prospectivity modeling and operational exploration targeting by converting statistical outputs into exploration recommendations.

1 Introduction

The depletion of near-surface deposits compels explorers to target deeper, concealed mineral systems with higher discovery costs [1, 2]. Mineral prospectivity mapping (MPM) now plays a central role in guiding exploration investments [3, 4]. Machine learning methods have transformed MPM [5, 6], evolving from logistic regression [7] to ensemble methods [5, 8, 12] and deep learning [9, 10, 11], with data augmentation techniques addressing class imbalance [13, 14].

However, three key problems remain. First, most studies treat all non-deposit locations as negatives, ignoring undiscovered deposits [15, 16, 17]; PU learning addresses this [18, 19] but remains underexplored in MPM [20, 21]. Second, spatial autocorrelation violates CV independence assumptions [22, 23, 24], with performance differences of up to 47% between spatial and non-spatial CV [25]. Third, models typically lack uncertainty measures essential for risk assessment [26, 27, 28, 29]; only 22.5% of Earth observation datasets incorporate uncertainty [33].

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.

We introduce a practical framework that tackles these problems. This paper makes three contributions. First, we develop a five-tier exploration priority classification that combines bootstrap probability, relative uncertainty, and conformal coverage to provide explicit action recommendations. Second, we implement a nested cross-validation design using spatial blocks sized according to variogram analysis, which prevents information leakage during hyperparameter optimization. Third, we adapt bagging-based PU learning [20] with XGBoost classifiers and systematically compare its performance against standard supervised learning.

We demonstrate our framework on the Yilgarn Craton, Western Australia, a world-class komatiite-hosted nickel province.

2 Related Work

2.1 Machine Learning for Mineral Prospectivity Mapping

Machine learning for MPM has evolved from logistic regression and weights-of-evidence [3, 7] to ensemble methods [5, 8, 12] and deep learning [9, 10, 11, 36]. Class imbalance remains a challenge addressed through data augmentation [13, 14] and SMOTE [37]. However, most studies treat all non-deposit locations as negative samples, conflicting with exploration reality.

2.2 Positive-Unlabeled Learning

PU learning addresses scenarios with only positive examples and unlabeled data [18, 19]. The SCAR assumption is problematic for mineral exploration due to spatial clustering and discovery bias [67, 68, 69]. Mordelet and Vert [20] proposed BaggingPU with increased robustness to SCAR violations [70]. MPM applications have grown recently [16, 17, 21], but few studies combine PU learning with spatial validation.

2.3 Spatial Cross-Validation

Standard k-fold CV assumes i.i.d. samples, violated by spatial autocorrelation [22]. Roberts et al. [22] recommended block CV exceeding the autocorrelation range, with Schratz et al. [25] reporting 47% performance differences between spatial and non-spatial CV. Cawley and Talbot [35] warned that spatially biased validation amplifies overfitting, motivating nested CV designs.

2.4 Conformal Prediction and Uncertainty Quantification

Uncertainty quantification transforms predictions into decision support [26, 28, 29]. Conformal prediction provides distribution-free coverage guarantees [31, 32], with extensions for spatial data [53, 71, 54]. Our spatial block design creates approximately independent splits for valid conformal inference.

Combining PU learning, spatial CV, and conformal prediction in a single MPM framework is the main contribution of this work.

3 Methodology

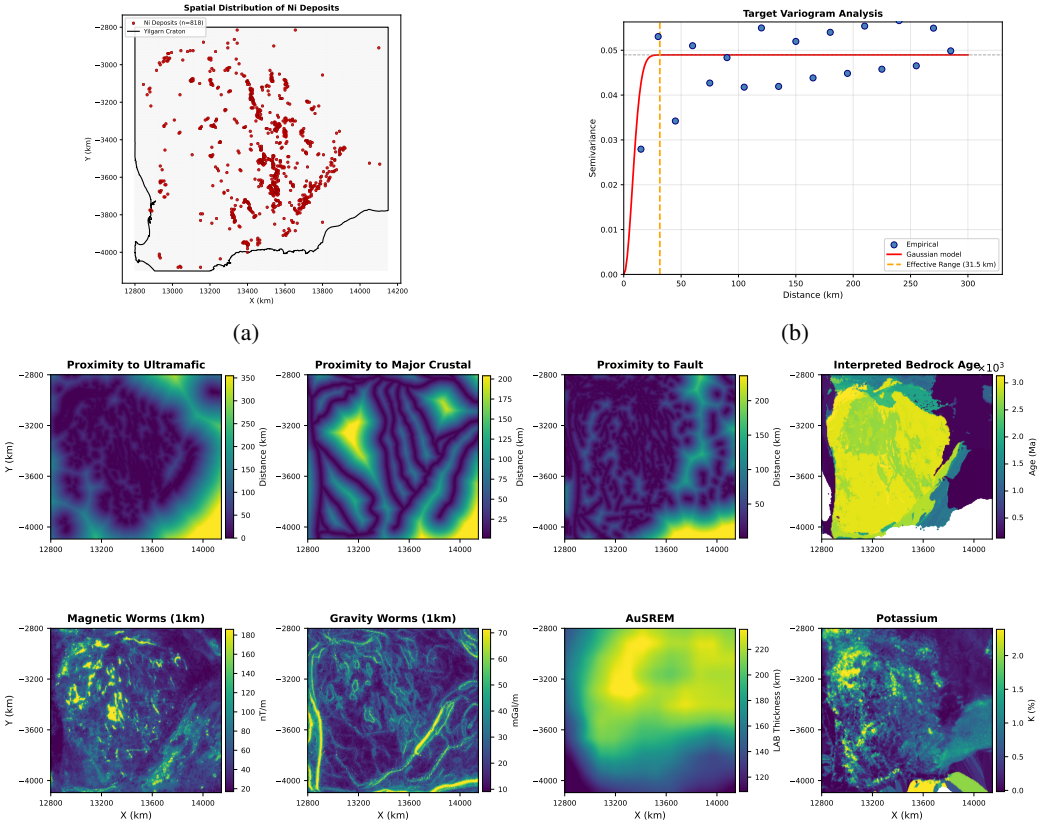
The following subsections detail the framework components.

3.1 Problem Formulation

We formalize MPM as a positive-unlabeled (PU) learning problem with positive samples $\mathcal{P} = \{x_i : y_i = 1\}$ (known deposits) and unlabeled samples $\mathcal{U} = \{x_j : y_j = ?\}$ containing both true negatives and undiscovered positives. Traditional binary classification treating unlabeled samples as negatives biases models against prospective regions [18]. Our framework combines BaggingPU-XGBoost, bootstrap uncertainty, cross-conformal prediction, and practical zone classification.

3.2 Yilgarn Craton Ni-Cu Dataset

We conduct experiments on the Yilgarn Craton, Western Australia, a 657,000 km^2 Archean granite-greenstone terrain hosting world-class komatiite-associated nickel deposits [55]. Figure 1 shows the study area and data compilation. We compile a GIS database from GSWA and GA including proximity features [57], geophysical grids [58], and multi-scale worm densities [59]. The dataset contains 70,200 grid cells (2 km resolution); positives are defined by proximity to known Ni deposit locations (818 cells, 1.17%) and the remaining 69,382 cells are unlabeled ($\sim$1:85 class imbalance). Table 1 summarizes 23 predictor features.



(a) (b) (c)

Figure 1: Study area, variogram analysis, and feature spatial distribution in the Yilgarn Craton, Western Australia. (a) Spatial distribution of 818 known Ni-Cu sulfide deposits overlaid on the regional geological map. (b) Experimental semivariogram fitted with a Gaussian model showing an effective range of 31.5 km. (c) Spatial distribution of eight representative predictor variables.

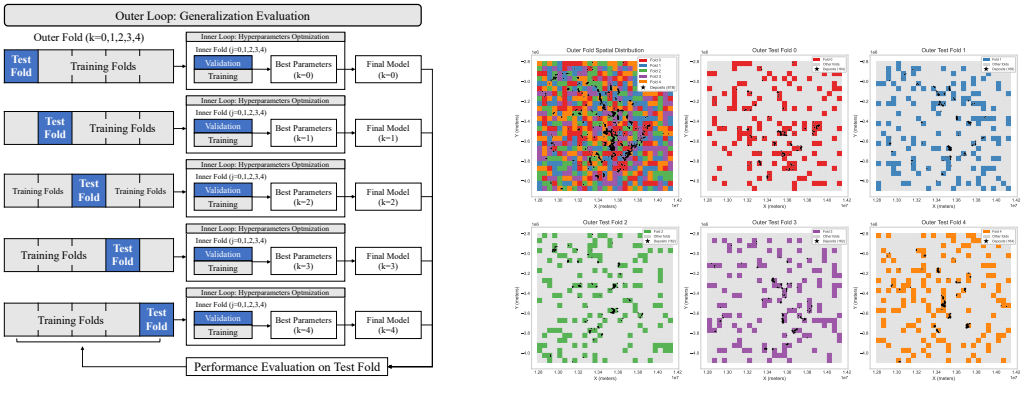
Table 1: Feature preprocessing summary.

Category	Features	Unit	Miss.(%)	Preprocessing
Proximity	Ultramafic, MajorCrustal, Fault	m	0	Log, RobustScaler
Geophysical	TMI, RTP, 1VD	nT, nT/m	6	Missing ind., Winsor., RobustScaler
Geophysical	DGIR, AuSREM	mGal, km	0	Winsor., RobustScaler
Worms	worms_mag (1, 4.6, 11, 50 km) worms_grav (1, 4.6, 11, 50 km)	nT/m, mGal/m	1	RobustScaler
Radiometric	K, Th, U	%, ppm	0	RobustScaler
Geological	Tectonic_Age, Interpreted_Age	Ma	0, 11	RobustScaler, Missing ind.

Preprocessing within CV folds uses training statistics only. Features with $>$5% missingness use missing indicators; magnetic/gravity features undergo Winsorization; proximity features are log-transformed; all features use RobustScaler normalization.

3.3 Nested Spatial Block Cross-Validation

Spatial autocorrelation violates CV independence assumptions, causing optimistic bias [22]. Variogram analysis (Figure 1b) yields an effective range of 31.5 km; we adopt 50 km spatial blocks (1.59 times the effective range) following Roberts et al. [22, 24, 25]. The 702 blocks are assigned to five stratified outer folds (162–166 positives each). Nested CV separates hyperparameter optimization (inner five-fold) from generalization assessment (outer test fold), preventing information leakage [35]. Figure 2 illustrates this structure.



(a) (b)

Figure 2: Nested spatial block cross-validation design. (a) Schematic of the nested structure with outer five-fold loop for generalization evaluation and inner five-fold loop for hyperparameter optimization. (b) Spatial distribution of five outer folds across the Yilgarn Craton, where each fold comprises spatially dispersed 50 km $\times$ 50 km blocks.

3.4 Evaluation Metrics

We adopt Precision-Recall Area Under Curve (PR-AUC) as the primary optimization metric for hyperparameter selection. For a classifier with varying threshold τ, precision and recall are computed as $\text{Precision}(\tau) = TP(\tau)/(TP(\tau) + FP(\tau))$ and $\text{Recall}(\tau) = TP(\tau)/(TP(\tau) + FN(\tau))$, with PR-AUC defined as the integral of precision over the recall range. PR-AUC is preferred over ROC-AUC for severely imbalanced datasets because it focuses on positive class performance without being influenced by the large number of true negatives. In our dataset with 1.17% positive rate, ROC-AUC can appear high even when the model fails to identify many deposits, whereas PR-AUC more sensitively reflects performance on the minority class of interest [60, 61]. We also report ROC-AUC as a traditional comparison metric.

3.5 Hyperparameter Optimization

We optimize XGBoost [8] hyperparameters using Optuna [62] with 30 trials per outer fold, PR-AUC objective on inner five-fold CV. The search space includes n_estimators [100, 500], max_depth [3, 10], learning_rate [0.01–0.1], min_child_weight [1, 15], gamma [0, 0.3], subsample [0.6–1.0], colsample_bytree [0.6–1.0], reg_alpha and reg_lambda [0, 1], and scale_pos_weight [1, 85]. BaggingPU [20] addresses PU learning by training base classifiers on bootstrap samples treating unlabeled data as negative, then aggregating predictions to reduce bias from mislabeled positives. Our BaggingPU-XGBoost uses two-stage optimization: Optuna for base XGBoost parameters, then grid search over n_estimators {10–50} and max_samples {0.1–0.5} (45 combinations). Complete hyperparameter configurations are in Appendix A.

3.6 Bootstrap Ensemble for Uncertainty Quantification

We employ bootstrap aggregation to estimate prediction uncertainty [30]. For each outer fold, we train 50 bootstrap models with optimized hyperparameters. For test sample x, we compute bootstrap mean probability $\bar{p}(x) = (1/50)\sum_{b=1}^{50} p_b(x)$ and quantify uncertainty using relative interquartile

range: rel_IQR$(x) = (Q_{75}(x) - Q_{25}(x))/\bar{p}(x)$. Our practical zone classification uses rel_IQR only for samples in the top 10% probability, where minimum thresholds (0.078 for BaggingPU, 0.022 for XGBoost) ensure numerical stability. The rel_IQR provides a scale-independent measure of prediction variability across regions with different prospectivity levels.

3.7 Cross-Conformal Calibration

Conformal prediction provides distribution-free prediction sets with guaranteed coverage [31, 32]. Our spatial block design (1.59 times the effective autocorrelation range) creates approximately independent splits satisfying approximate exchangeability conditions for valid conformal inference [71, 53]. For each outer fold, we apply 20-iteration cross-conformal calibration:

1. **Probability calibration**: Platt scaling [63]: $p_{\text{cal}}(x) = 1/(1 + \exp(A \cdot \bar{p}(x) + B))$.
2. **Threshold computation**: For target FNR control level $\alpha = 0.15$, compute threshold τ such that FNR$(\tau) \leq \alpha$.
3. **Prediction set construction**: Assign in_set_1$(x) = \mathbf{1}[p_{\text{cal}}(x) \geq \tau]$.

The procedure guarantees $\geq$85% true positive coverage. Empirically, we achieve 96.08% (XGBoost) and 96.20% (BaggingPU) coverage, exceeding the nominal guarantee due to Hoeffding bound correction for finite sample size ($\sim$11.9% with $\sim$131 positives per calibration fold), median aggregation across 20 cross-conformal iterations, and spatial block separation ensuring conservative thresholds. This over-coverage is desirable in MPM where false negative costs (missed deposits) exceed false positive costs.

3.8 Practical Zone Classification

The practical zone classification system uses probability, uncertainty, and conformal membership to rank exploration targets. Figure 3 illustrates the workflow combining bootstrap mean probability, rel_IQR uncertainty, and conformal membership to assign each grid cell to one of five exploration priority zones.

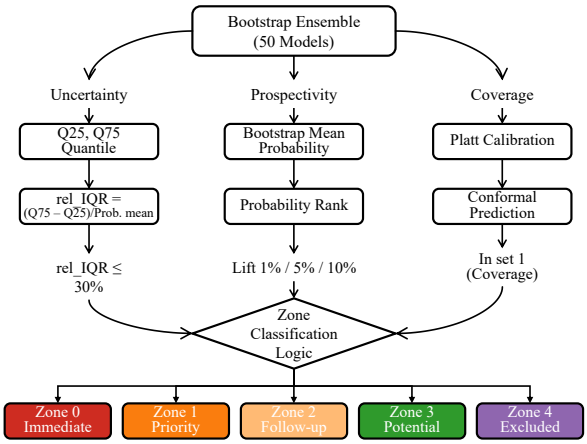


Figure 3: Practical zone classification framework integrating bootstrap ensemble predictions, relative IQR uncertainty (30% threshold), and conformal coverage for five-tier zone assignment. Samples failing conformal coverage are assigned to Zone 4 (EXCLUDED) regardless of probability.

We define five exploration priority zones based on a 30% rel_IQR threshold distinguishing low and high uncertainty. Zone 0 (IMMEDIATE) requires top 1% probability with rel_IQR $\leq$ 30% for immediate drilling. Zone 1 (PRIORITY) includes top 1% with high uncertainty (rel_IQR $>$ 30%) or top 1–5% with low uncertainty (rel_IQR $\leq$ 30%). Zone 2 (FOLLOW-UP) covers top 1–5% with high uncertainty or top 5–10% with low uncertainty. Zone 3 (POTENTIAL) includes remaining samples with conformal coverage. Zone 4 (EXCLUDED) contains samples failing conformal coverage (in_set_1 = 0).

Table 2: Prediction performance comparison. Inner CV shows hyperparameter optimization phase; Outer Test shows generalization evaluation on spatially independent regions.

	Inner CV Performance				Outer Test Performance			
Fold	XGB PR	Bag PR	XGB ROC	Bag ROC	XGB PR	Bag PR	XGB ROC	Bag ROC
0	0.169	0.187	0.900	0.912	0.109	0.106	0.874	0.875
1	0.158	0.190	0.888	0.897	0.201	0.196	0.940	0.940
2	0.153	0.160	0.898	0.908	0.204	0.195	0.912	0.911
3	0.169	0.182	0.908	0.917	0.150	0.152	0.886	0.886
4	0.135	0.143	0.882	0.893	0.313	0.322	0.944	0.945
Mean	0.157	0.172	0.895	0.905	0.195	0.194	0.911	0.911
Std	0.013	0.019	0.009	0.009	0.076	0.080	0.028	0.028

XGB = XGBoost, Bag = BaggingPU-XGBoost, PR = PR-AUC, ROC = ROC-AUC

The lift-based approach using relative percentiles keeps zone definitions consistent across models. The 30% rel_IQR threshold, selected empirically (Section 4.4), provides a practical reference conceptually inspired by uncertainty conventions in mineral resource estimation [34]. The zone classification uses bootstrap mean probability (mineralization potential), rel_IQR uncertainty (prediction reliability), and conformal membership (exclusion filter). We compare XGBoost and BaggingPU-XGBoost, focusing on deposit capture in high-priority zones (0–2) while minimizing Zone 4 exclusions.

4 Experiments and Analysis

4.1 Experimental Setup

Experiments were conducted on Intel Core i9-14900K CPU (3.20 GHz), 64 GB RAM, and NVIDIA RTX 4090 GPU, with Python 3.12, XGBoost 2.0.3, scikit-learn 1.4.0, Optuna 3.5.0, and NumPy 1.26.4. Fixed seed (42) ensures reproducibility. For each outer fold: 30 Optuna trials optimize XGBoost hyperparameters on inner CV (PR-AUC objective); BaggingPU uses grid search over 45 parameter combinations; 50 bootstrap models provide uncertainty estimates; 20-iteration cross-conformal calibration targets 85% coverage ($\alpha = 0.15$). Code will be released upon publication.

4.2 Prediction Performance Analysis

We summarize the model performance comparison in Table 2, which presents inner CV and outer test performance for both models. Inner CV metrics represent mean performance across five inner folds (optimization signal); outer test metrics represent generalization to spatially independent held-out regions.

BaggingPU-XGBoost achieves higher inner CV PR-AUC (0.172 vs. 0.157), reflecting reduced bias from potential undiscovered positives [20]. On outer test folds, both models converge to comparable mean PR-AUC ($\sim$0.195) and identical ROC-AUC (0.911), suggesting that PU learning benefits observed during training do not directly translate to aggregate performance gains in spatially independent regions. This may reflect the Yilgarn Craton's maturity as an exploration province with fewer undiscovered deposits [72].

The substantial fold-level variation (PR-AUC 0.106–0.322) reflects fundamental differences in mineralization predictability across spatial blocks, arising from how well predictor-response relationships transfer to held-out blocks. This spatial heterogeneity, formalized by Meyer and Pebesma [42] as "area of applicability," presents a major challenge for MPM in underexplored regions. The fold-level variability (std 0.07–0.08 for outer test vs. 0.01–0.02 for inner CV) quantifies additional uncertainty when predicting in novel spatial contexts—the greenfield exploration scenario. Our nested CV design separates hyperparameter optimization from generalization assessment, preventing the optimistic bias that standard CV produces [35].

4.3 Conformal Calibration Analysis

We present the cross-conformal calibration analysis in Table 3, which summarizes the calibration results.

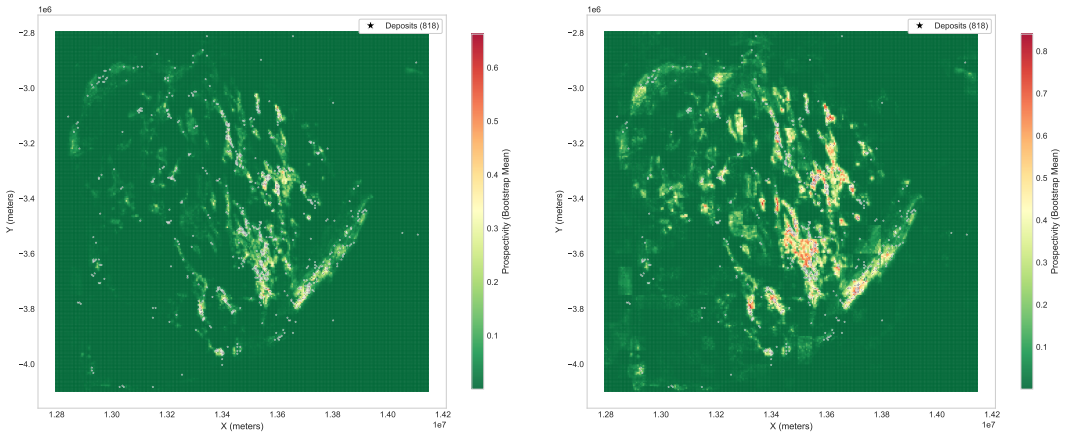
Table 3: Cross-conformal calibration analysis showing Platt scaling parameters and coverage statistics across five outer folds.

Parameter	XGBoost	BaggingPU
Platt Coefficient (mean $\pm$ std)	7.54 ± 0.43	6.74 ± 1.07
Platt Intercept (mean $\pm$ std)	-4.59 ± 0.02	-4.97 ± 0.26
FNR Threshold (mean $\pm$ std)	0.0102 ± 0.0002	0.0071 ± 0.0017
Coverage (mean $\pm$ std)	$96.08\% \pm 1.44\%$	$96.20\% \pm 2.17\%$
FNR (mean $\pm$ std)	$3.92\% \pm 1.44\%$	$3.80\% \pm 2.17\%$

Both models substantially exceed the nominal 85% coverage, achieving 96.08% (XGBoost) and 96.20% (BaggingPU) with corresponding FNR of 3.92% and 3.80%. This conservative behavior is desirable in exploration where false negative costs exceed false positive costs. XGBoost requires a steeper Platt calibration curve (7.54 vs. 6.74), indicating narrower raw probability dynamic range. The low FNR thresholds (0.0102 and 0.0071) are consistent with severe class imbalance.

4.4 Practical Zone Analysis

The practical zone framework classifies samples into five exploration tiers. The spatial prospectivity maps showing bootstrap mean probability predictions are visualized in Figure 4. Zone statistics using lift-based classification are summarized in Table 4.



(a) (b)

Figure 4: Spatial prospectivity maps showing bootstrap mean probability predictions. (a) XGBoost; (b) BaggingPU-XGBoost. Both models identify similar high-prospectivity regions around known ultramafic complexes.

The key difference appears in Zone 0 distribution. BaggingPU-XGBoost assigns 0.68% of samples to Zone 0 capturing 17.25% of deposits, while XGBoost assigns only 0.02% capturing 0.73%. This arises because BaggingPU produces substantially lower rel_IQR in high-probability regions (median 0.18 vs. 0.48 in top 1%), giving more stable predictions in high-probability regions. This robustness to SCAR assumption violations [67, 68] makes BaggingPU a particularly suitable approach for MPM [69, 70]. XGBoost shows higher capture in Zones 1–2 individually due to redistribution from elevated rel_IQR values.

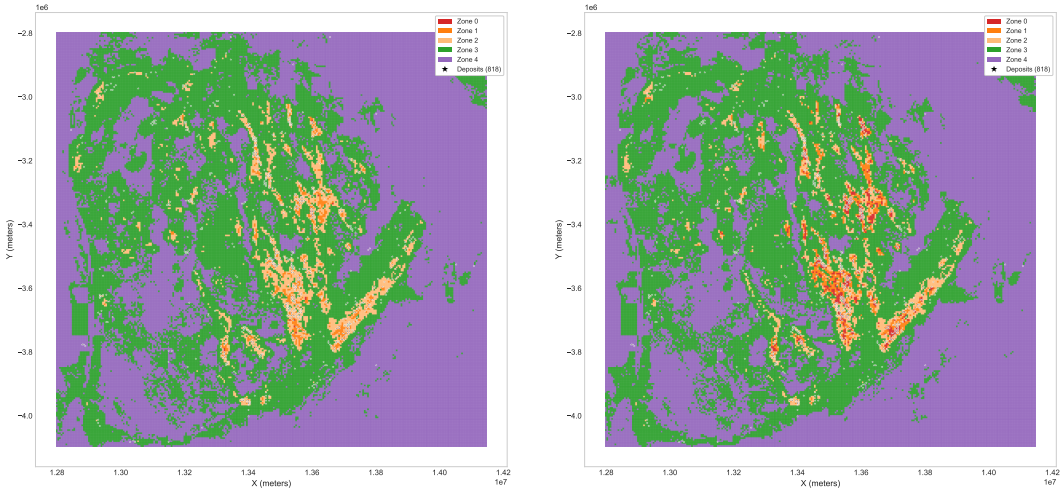
Despite these differences, both models achieve nearly identical cumulative capture for Zones 0–2: $\sim$55% of deposits within $\sim$5% of study area, an 11-fold concentration factor [74]. Both maintain low Zone 4 deposit loss ($<$4%), ensuring conformal exclusion removes minimal deposits from

Table 4: Practical zone capture statistics (LIFT mode). Zone-level deposit capture rate and area percentage with cumulative capture for high-priority zones (0–2).

Zone	BaggingPU Capture	BaggingPU Area	XGBoost Capture	XGBoost Area
0 (IMMEDIATE)	$17.25\% \pm 10.73\%$	0.68%	$0.73\% \pm 1.10\%$	0.02%
1 (PRIORITY)	$15.54\% \pm 7.19\%$	1.38%	$21.88\% \pm 6.57\%$	0.98%
2 (FOLLOW-UP)	$22.19\% \pm 10.34\%$	3.02%	$32.24\% \pm 6.83\%$	4.00%
3 (POTENTIAL)	$41.22\% \pm 11.13\%$	39.75%	$41.22\% \pm 10.76\%$	39.57%
4 (EXCLUDED)	$3.80\% \pm 2.17\%$	55.17%	$3.92\% \pm 1.44\%$	55.42%
Cumulative (0–2)	**54.98%**	**5.08%**	**54.86%**	**5.00%**

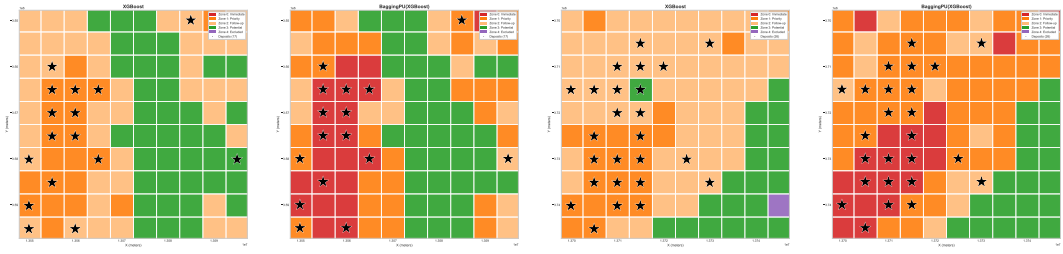
consideration. With either model, targeting Zones 0–2 would examine ∼33,000 km^2 (5% of 657,000 km^2 Yilgarn Craton) while capturing over half of known deposits.

The spatial distribution of practical zones is presented in Figure 5, and detailed block-level examples from Fold 2 and Fold 4 test regions are provided in Figure 6.



(a) (b)

Figure 5: Spatial practical zone classification maps under the lift-based framework. (a) XGBoost; (b) BaggingPU-XGBoost. The five-tier classification ranges from Zone 0 (IMMEDIATE) to Zone 4 (EXCLUDED), with high-priority zones concentrated around known ultramafic complexes.



(a) (b) (c) (d)

Figure 6: Detailed block-level practical zone analysis for representative spatial blocks. (a, b) Block from Fold 2: (a) XGBoost, (b) BaggingPU-XGBoost. (c, d) Block from Fold 4: (c) XGBoost, (d) BaggingPU-XGBoost.

TreeSHAP feature importance analysis (Appendix C) confirms that both models identify proximity to ultramafic sources as the dominant predictor, with the top five features remaining consistent across

XGBoost and BaggingPU-XGBoost, showing that PU learning preserves interpretable geological relationships.

4.5 Limitations

This study has several limitations. First, bootstrap uncertainty may underestimate total predictive uncertainty, particularly for predictions distant from training data domains [75], and does not formally distinguish between aleatoric and epistemic uncertainty components as Bayesian approaches could provide [76, 29]. Second, while our spatial block design creates approximately independent splits for valid conformal inference [71, 53], coverage guarantees should be interpreted as approximate for strongly autocorrelated geological settings; recent developments in GeoConformal Prediction [54] offer promising directions for strengthening spatial coverage guarantees. Third, the fixed 50 km isotropic block size may not optimally capture anisotropic spatial dependencies present in geological features; adaptive blocking strategies [44, 45] warrant investigation. Finally, this study focuses on a single geological province (Yilgarn Craton) and commodity type (Ni); predictor-response relationships may not transfer directly to other tectonic settings or mineralization styles [72]. Future studies should test the framework on other regions and improve uncertainty estimation.

5 Conclusion

This paper presented a framework for mineral prospectivity mapping combining uncertainty quantification with conformal coverage guarantees for exploration targeting. The framework handles label ambiguity with PU learning, spatial autocorrelation with nested block CV, and decision support with zone classification.

The practical zone classification synthesizes bootstrap probability, rel_IQR uncertainty (30% threshold), and conformal membership to assign spatial units to five exploration tiers. Validation on the Yilgarn Craton demonstrates 55% deposit capture within 5% of study area (11-fold concentration), with conformal exclusion removing less than 4% of deposits. BaggingPU-XGBoost produces more stable high-probability predictions (median rel_IQR 0.18 vs. 0.48), enabling confident identification of immediate targets.

Our framework bridges the gap between academic prospectivity modeling and operational targeting by converting predictions into ranked exploration priorities. The method can apply to other commodities and regions, but zone thresholds need adjustment based on local conditions. Multi-region validation and improved uncertainty methods remain as future work.

References

[1] Schodde, R. (2017). Long term outlook for the global exploration industry. MinEx Consulting.

[2] Davies, R.S., et al. (2021). Learning and Expertise in Mineral Exploration Decision-Making. *International Journal of Environmental Research and Public Health*, 18(18), 9752.

[3] Carranza, E.J.M. (2008). Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. Elsevier.

[4] Zuo, R. & Carranza, E.J.M. (2011). Support vector machine: A tool for mapping mineral prospectivity. *Computers & Geosciences*, 37(12), 1967–1975.

[5] Rodriguez-Galiano, V.F., et al. (2015). Machine learning predictive models for mineral prospectivity. *Ore Geology Reviews*, 71, 804–818.

[6] Zuo, R. & Carranza, E.J.M. (2023). Machine Learning-Based Mapping for Mineral Exploration. *Mathematical Geosciences*, 55, 891–895.

[7] Bonham-Carter, G.F. (1994). Geographic Information Systems for Geoscientists. Pergamon Press.

[8] Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794.

[9] Li, Q., Chen, G. & Luo, L. (2023). Mineral Prospectivity Mapping Using Attention-Based CNN. *Ore Geology Reviews*, 156, 105381.

[10] Xu, Y. & Zuo, R. (2024). An Interpretable Graph Attention Network for MPM. *Mathematical Geosciences*, 56, 169–190.

[11] Dong, Y.-L. & Zhang, Z.-J. (2024). Deep Forest Modeling: An Interpretable Deep Learning Method for Mineral Prospectivity Mapping. *Journal of Geophysical Research: Machine Learning and Computation*, 1(4), e2024JH000311.

[12] Yin, J. & Li, N. (2022). Ensemble Learning with Bayesian Optimization for MPM. *Ore Geology Reviews*, 145, 104916.

[13] Yang, N., et al. (2022). Data Augmentation in MPM Based on CNN. *Computers & Geosciences*, 161, 105075.

[14] Mantilla-Dulcey, A., et al. (2024). Porphyry-Type MPM with Transfer Learning. *Gondwana Research*, 136, 236–250.

[15] Sun, K., et al. (2024). A Review of MPM Using Deep Learning. *Minerals*, 14(10), 1021.

[16] Carranza, E.J.M. (2011). Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. *Journal of Geochemical Exploration*, 110(2), 167–185.

[17] Xiong, Y. & Zuo, R. (2021). A positive and unlabeled learning algorithm for mineral prospectivity mapping. *Computers & Geosciences*, 147, 104667.

[18] Elkan, C. & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, 213–220.

[19] Bekker, J. & Davis, J. (2020). Learning from positive and unlabeled data: a survey. *Machine Learning*, 109, 719–760.

[20] Mordelet, F. & Vert, J.-P. (2014). A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37, 201–209.

[21] Zhang, S., et al. (2022). An Integrated Framework for MPM Using Bagging-Based PU Learning. *Natural Resources Research*, 31, 3041–3060.

[22] Roberts, D.R., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929.

[23] Ploton, P., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11, 4540.

[24] Valavi, R., et al. (2019). blockCV: An R package for spatially separated folds for k-fold CV. *Methods in Ecology and Evolution*, 10(2), 225–232.

[25] Schratz, P., et al. (2019). Hyperparameter tuning and performance assessment using spatial data. *Ecological Modelling*, 406, 109–120.

[26] Lindi, O.T., et al. (2024). Uncertainty Quantification in Mineral Resource Estimation. *Natural Resources Research*, 33(6), 2503–2526.

[27] Olierook, H.K.H., et al. (2021). Bayesian Geological and Geophysical Data Fusion for 3D Geological Models. *Geoscience Frontiers*, 12(1), 479–493.

[28] Zhang, Z., et al. (2024). An Uncertainty-Quantification ML Framework for 3D MPM. *Natural Resources Research*, 33, 1393–1411.

[29] Liu, Y., et al. (2025). Dirichlet-Based Uncertainty-Aware Deep Learning for Explainable MPM. *Natural Resources Research*.

[30] Efron, B. & Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall.

[31] Vovk, V., Gammerman, A. & Shafer, G. (2005). Algorithmic Learning in a Random World. Springer.

[32] Angelopoulos, A.N. & Bates, S. (2021). A Gentle Introduction to Conformal Prediction. arXiv:2107.07511.

[33] Singh, G., Moncrieff, G., Venter, Z., et al. (2024). Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. *Scientific Reports*, 14, 16166.

[34] JORC (2012). Australasian Code for Reporting of Exploration Results, Mineral Resources and Ore Reserves.

[35] Cawley, G.C. & Talbot, N.L.C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias. *Journal of Machine Learning Research*, 11, 2079–2107.

[36] McMillan, M., et al. (2021). Mineral Prospectivity Mapping Using a VNet CNN. *The Leading Edge*, 40(2), 99–105.

[37] Li, B., Yu, Z. & Ke, X. (2023). One-Dimensional CNN for Mapping Mineral Prospectivity. *Ore Geology Reviews*, 160, 105573.

[38] du Plessis, M.C., Niu, G. & Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*.

[39] Kiryo, R., et al. (2017). Positive-Unlabeled Learning with Non-Negative Risk Estimator. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

[40] Kumar, P. & Lambert, C.G. (2024). PULSNAR: class proportion estimation when SCAR does not hold. *PeerJ Computer Science*, 10, e2451.

[41] Teisseyre, P., et al. (2024). Verifying the SCAR Assumption in PU Learning. arXiv:2404.00145.

[42] Meyer, H. & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability. *Methods in Ecology and Evolution*, 12(9), 1620–1633.

[43] Mila, C., et al. (2022). Nearest Neighbour Distance Matching Leave-One-Out CV for map validation. *Methods in Ecology and Evolution*, 13, 1304–1316.

[44] Wang, J., et al. (2023). Spatial+: A new CV method to evaluate geospatial ML models. *International Journal of Applied Earth Observation and Geoinformation*, 121, 103364.

[45] Linnenbrink, J., et al. (2024). kNNDM CV: k-fold nearest-neighbour distance matching CV. *Geoscientific Model Development*, 17, 5897–5912.

[46] Tziachris, P., et al. (2023). Spatial or Random Cross-Validation? The Effect of Resampling Methods in Predicting Groundwater Salinity with Machine Learning in Mediterranean Region. *Water*, 15(12), 2278.

[47] Wadoux, A.M.J.-C., et al. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457, 109692.

[48] Mery, N. & Marcotte, D. (2022). Quantifying Mineral Resources and Their Uncertainty Using ML Methods. *Mathematical Geosciences*, 54, 363–387.

[49] Fontana, M., Zeni, G. & Vantini, S. (2023). Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 29(1), 1–23.

[50] Angelopoulos, A.N., et al. (2024). Conformal Risk Control. *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

[51] Barber, R.F., et al. (2021). Predictive Inference with the Jackknife+. *Annals of Statistics*, 49(1), 486–507.

[52] Romano, Y., Patterson, E. & Candes, E. (2019). Conformalized Quantile Regression. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

[53] Barber, R.F., et al. (2023). Conformal Prediction Beyond Exchangeability. *Annals of Statistics*, 51(2), 816–845.

[54] Lou, X., Luo, P. & Meng, L. (2025). GeoConformal Prediction: A Model-Agnostic Framework for Measuring the Uncertainty of Spatial Prediction. *Annals of the American Association of Geographers*, 115(8), 1971–1998.

[55] Barnes, S.J., Fiorentini, M.L. & Fardon, M.C. (2012). Platinum group element and nickel sulphide ore tenors of the Mount Keith nickel deposit. *Mineralium Deposita*, 47(1-2), 129–150.

[56] Cassidy, K.F., et al. (2006). A revised geological framework for the Yilgarn Craton. Record 2006/8, Geological Survey of Western Australia.

[57] Begg, G.C., et al. (2010). Lithospheric, cratonic, and geodynamic setting of Ni-Cu-PGE sulfide deposits. *Economic Geology*, 105(6), 1057–1070.

[58] Kennett, B.L.N. & Salmon, M. (2012). AuSREM: Australian Seismological Reference Model. *Australian Journal of Earth Sciences*, 59(8), 1091–1103.

[59] Holden, E.-J., Dentith, M. & Kovesi, P. (2008). Towards the automated analysis of regional aeromagnetic data to identify regions prospective for gold deposits. *Computers & Geosciences*, 34(11), 1505–1513.

[60] Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, 233–240.

[61] Sofaer, H.R., Hoeting, J.A. & Jarnevich, C.S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4), 565–577.

[62] Akiba, T., et al. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, 2623–2631.

[63] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61–74.

[64] Carranza, E.J.M. (2009). Objective selection of suitable unit cell size in data-driven modeling of mineral prospectivity. *Computers & Geosciences*, 35(10), 2032–2046.

[65] Porwal, A. & Carranza, E.J.M. (2015). Introduction to the Special Issue: GIS-based mineral potential modelling and geological data analyses for mineral exploration. *Ore Geology Reviews*, 71, 477–483.

[66] Singer, D.A. & Kouda, R. (1999). A comparison of the weights-of-evidence method and probabilistic neural networks. *Natural Resources Research*, 8(4), 287–298.

[67] Lisitsin, V., Porwal, A. & McCuaig, T.C. (2015). Spatial data analysis of mineral deposit point patterns: Applications to exploration targeting. *Ore Geology Reviews*, 71, 861–881.

[68] Coolbaugh, M.F., Raines, G.L. & Zehner, R.E. (2007). Assessment of exploration bias in data-driven predictive models and the estimation of undiscovered resources. *Natural Resources Research*, 16(2), 199–207.

[69] Zhang, S., et al. (2025). Recursive Annotation for Negative Labeling in Data-Driven Mineral Prospectivity Mapping. *Natural Resources Research*, 34, 2373–2402.

[70] Claesen, M., De Smet, F., Suykens, J.A.K. & De Moor, B. (2015). A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing*, 160, 73–84.

[71] Mao, H., Martin, R. & Reich, B.J. (2024). Valid model-free spatial prediction. *Journal of the American Statistical Association*, 119(546), 904–914.

[72] Ludwig, M., et al. (2023). Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography*, 32(3), 356–368.

[73] Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A. & Burnaev, E. (2024). Challenges in data-driven geospatial modeling for environmental research and practice. *Nature Communications*, 15, 10700.

[74] Yousefi, M. & Carranza, E.J.M. (2015). Prediction-area (P-A) plot and C-A fractal analysis to classify and evaluate evidential maps for mineral prospectivity modeling. *Computers & Geosciences*, 79, 69–81.

[75] Palmer, G., Du, S., Politowicz, A., Emory, J.P., Yang, X., Gautam, A. & Morgan, D. (2022). Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Computational Materials*, 8(1), 115.

[76] Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

[77] Lundberg, S.M. & Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

[78] Yao, J., Mao, X., Zhou, Y., et al. (2024). Mineral prospectivity mapping susceptibility evaluation based on interpretable ensemble learning. *Ore Geology Reviews*, 172, 106248.

[79] Li, Z., Chen, G., Ma, L., et al. (2024). Optimization of Feature Selection in Mineral Prospectivity Mapping. *Minerals*, 14(10), 970.

[80] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

A Hyperparameter Optimization Details

We optimized XGBoost hyperparameters using Optuna [62] with 30 trials per outer fold (PR-AUC objective on inner CV). Table 5 presents the best (Max) and worst (Min) configurations from each fold.

Table 5: XGBoost hyperparameters from Optuna optimization (30 trials per fold), showing best (Max) and worst (Min) PR AUC configurations.

Fold	PR AUC	n_est	max_depth	lr	mcw	gamma	reg_alpha	reg_lambda
0	0.169	308	7	0.023	10	0.100	1.66e-03	1.09e-07
	0.131	135	4	0.012	4	0.194	2.77e-06	0.288
1	0.158	222	5	0.029	4	0.291	2.99e-05	1.05e-05
	0.119	135	4	0.012	4	0.194	2.77e-06	0.288
2	0.153	243	4	0.063	2	0.401	4.69e-08	7.62
	0.121	289	6	0.270	3	0.342	2.23e-08	0.037
3	0.168	279	6	0.029	9	0.437	8.16e-07	2.50e-06
	0.122	447	6	0.111	1	0.485	0.310	8.15e-07
4	0.135	243	4	0.063	2	0.401	4.69e-08	7.62
	0.104	289	6	0.270	3	0.342	2.23e-08	0.037

For each fold, the first row shows Max (best) and the second row shows Min (worst). lr = learning_rate, mcw = min_child_weight

The optimization achieved mean PR AUC of 0.157 ± 0.013 across folds; Folds 2 and 4 converged to identical configurations due to Optuna's TPE sampler generating identical trial sequences with the same random seed [62]. Best-performing configurations exhibited lower learning rates (0.023–0.063) versus worst performers (0.012–0.270), consistent with Friedman's [80] finding that smaller learning rates produce more robust models. The subsample and colsample_bytree parameters were fixed at 1.0 to isolate the PU learning effect for fair BaggingPU comparison.

Table 6: BaggingPU bagging parameters from GridSearch (45 combinations per fold), showing best (Max) and worst (Min) PR AUC configurations.

Fold	PR AUC	n_estimators	max_samples
0	0.187	20	0.50
	0.180	20	0.15
1	0.190	20	0.40
	0.180	10	0.15
2	0.160	30	0.10
	0.149	40	0.45
3	0.182	40	0.50
	0.175	10	0.20
4	0.143	10	0.10
	0.134	10	0.50

For each fold, the first row shows Max (best) and the second row shows Min (worst).

For BaggingPU-XGBoost, we employed grid search over n_estimators {10–50} and max_samples {0.1–0.5} (45 combinations), using Optuna-optimized XGBoost as the base estimator. The optimization achieved mean PR AUC of 0.172 ± 0.019. Optimal max_samples showed spatial partition-dependent patterns: Folds 0, 1, 3 preferred larger values (0.40–0.50), while Folds 2, 4 preferred smaller values (0.10), consistent with Mordelet and Vert's [20] observation that optimal sampling rate varies with hidden positive contamination in the unlabeled set.

B Probability-Based Zone Classification Alternative

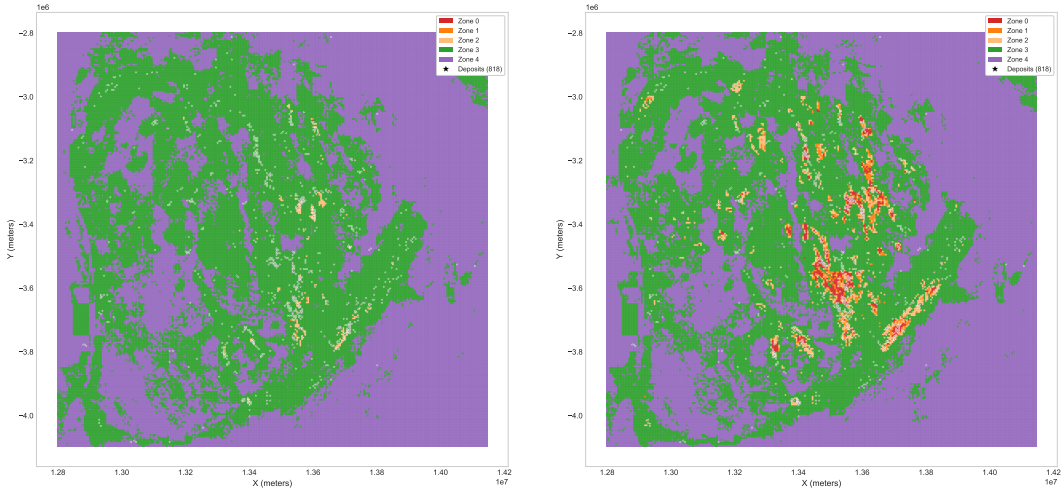
This appendix describes an alternative zone classification approach based on absolute probability thresholds. The probability-based mode (Prob Mode) follows the same framework as the lift-based approach presented in Section 3.8, but replaces the percentile-based thresholds (top 1%, 5%, 10%)

with fixed probability thresholds (0.50, 0.25, 0.10). All other criteria, the rel_IQR $\leq 30\%$ uncertainty threshold and conformal coverage requirement, remain unchanged. Table 7 compares the deposit capture rates between Prob Mode and Lift Mode for both models. The comparison reveals substantial differences in zone assignment effectiveness.

Table 7: Deposit capture rate comparison between Prob Mode and Lift Mode (mean $\pm$ std across five folds).

Model	Mode	Zone 0+1 Capture	Zone 0–2 Capture	Zone 4 Miss
XGBoost	Prob	$0.86\% \pm 1.03\%$	$10.29\% \pm 5.05\%$	$3.92\% \pm 1.44\%$
XGBoost	Lift	$22.62\% \pm 6.16\%$	$54.86\% \pm 11.56\%$	$3.92\% \pm 1.44\%$
BaggingPU	Prob	$28.15\% \pm 23.42\%$	$41.61\% \pm 24.17\%$	$3.80\% \pm 2.17\%$
BaggingPU	Lift	$32.78\% \pm 17.33\%$	$54.98\% \pm 12.46\%$	$3.80\% \pm 2.17\%$

Lift Mode substantially outperforms Prob Mode for both models, particularly for XGBoost where Zone 0–2 capture increases from 10.29% to 54.86%. The Zone 4 miss rates remain identical because conformal coverage is independent of the probability threshold scheme. The superior performance of lift-based classification stems from its invariance to model calibration: in severely imbalanced datasets (1.17% positive rate), the 0.50 probability threshold for Zone 0 under Prob Mode is rarely achieved, resulting in minimal high-priority zone assignments. Lift-based thresholds automatically adapt to the probability distribution, ensuring meaningful zone assignments regardless of calibration. However, Prob Mode may be preferred when absolute probability interpretability is important or when comparing results across different study areas. Figure 7 presents spatial prospectivity maps under Prob Mode for visual comparison with the Lift Mode results shown in Section 4.4. The contrast between the two classification approaches is visually striking: XGBoost under Prob Mode produces almost no Zone 0 or Zone 1 assignments, while BaggingPU shows moderate differentiation due to its higher predicted probabilities for prospective areas.



(a) (b)

Figure 7: Spatial practical zone maps under probability-based (Prob Mode) classification. (a) XGBoost; (b) BaggingPU-XGBoost.

C TreeSHAP Feature Importance Analysis

We computed TreeSHAP values [77] across all five outer folds using 50 bootstrap models per fold (250 total evaluations, 10,000 test samples) to examine how positive-unlabeled learning affects feature attribution. For BaggingPU models, SHAP values were averaged across the 20 internal XGBoost base estimators. Figure 8 presents the complete feature importance comparison, excluding five missing indicator features with zero importance.

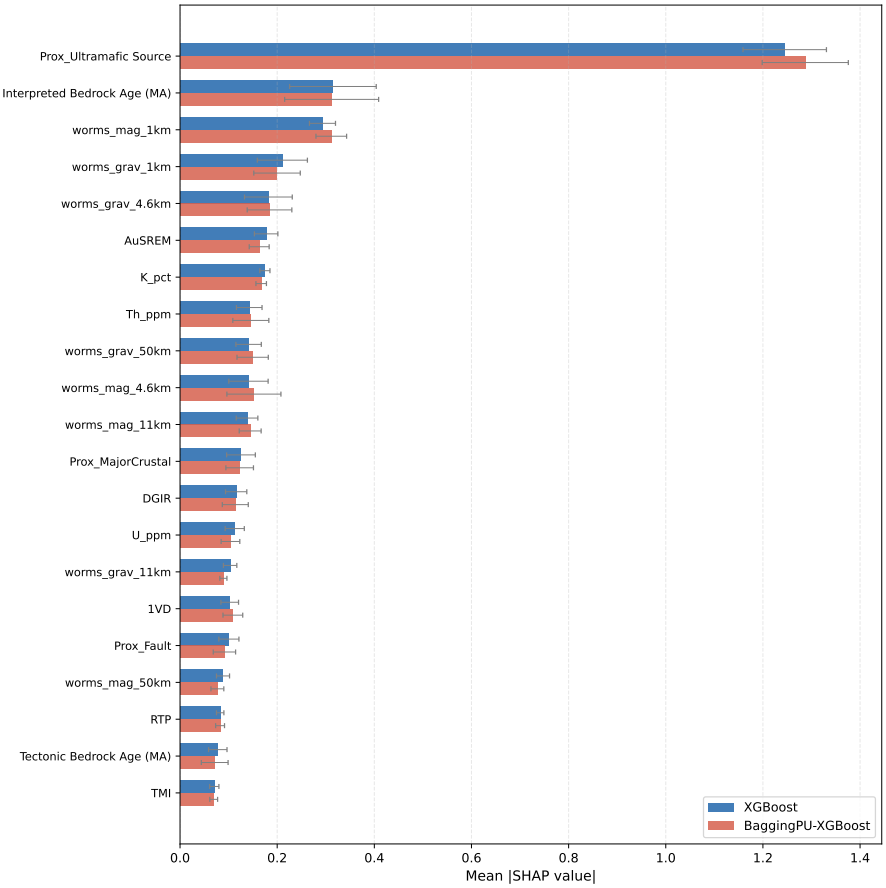


Figure 8: Comparison of mean |SHAP| values for all features between XGBoost and BaggingPU-XGBoost models averaged across 250 model evaluations (5 folds $\times$ 50 bootstraps).

The analysis reveals remarkable consistency between models. Proximity to ultramafic sources emerges as the dominant predictor with mean |SHAP| values of 1.244 (XGBoost) and 1.286 (BaggingPU), approximately four times higher than the second-ranked feature—consistent with established geological controls for Yilgarn Craton nickel sulfide deposits [55, 57]. The top five features remain identical across both models (Proximity to Ultramafic Source, Interpreted Bedrock Age, magnetic and gravity worm densities), with only minor rank exchanges between positions 2–3. No feature exhibited a rank change exceeding two positions, demonstrating that BaggingPU preserves interpretable geological relationships while providing uncertainty quantification benefits.

AI Co-Scientist Challenge Korea Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state three contributions: (1) Practical Zone Framework integrating bootstrap mean probability, relative uncertainty (30% rel_IQR threshold conceptually inspired by resource estimation conventions), and conformal coverage into five exploration priority tiers, achieving 55% deposit capture in 5% area (11-fold concentration factor); (2) Nested Spatial Block CV (outer five-fold $\times$ inner five-fold, 50 km blocks based on variogram analysis with 31.5 km effective range) for unbiased generalization assessment; and (3) BaggingPU-XGBoost for positive-unlabeled learning addressing label ambiguity inherent in MPM. All claims are supported by experimental results in Section 4 with quantitative evidence including PR-AUC (0.195), ROC-AUC (0.911), and 96% conformal coverage.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4.5 comprehensively discusses limitations including: (1) bootstrap-based uncertainty quantification captures ensemble disagreement but may underestimate total predictive uncertainty, particularly for predictions far from training data domains, and does not formally distinguish aleatoric from epistemic components as Bayesian methods could; (2) conformal prediction coverage guarantees should be interpreted as approximate rather than exact for strongly autocorrelated geological settings, though spatial block design creates approximately independent calibration-test splits; (3) fixed 50 km isotropic block size may not optimally capture anisotropic spatial dependencies in geological features such as fault systems; and (4) single geological province (Yilgarn Craton) and commodity type (Ni), with transferability to other tectonic settings or mineralization styles requiring validation. Future research directions are provided.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper is primarily empirical and does not introduce novel theoretical results requiring formal proofs. The conformal prediction guarantees cited are based on established theory from Vovk et al. [31] and Angelopoulos and Bates [32]. The exchangeability assumption for conformal prediction is explicitly acknowledged in Section 3.7, noting that spatial block design satisfies approximate exchangeability conditions per Mao et al. [71] and Barber et al. [53].

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3.5 provides complete hyperparameter search spaces and optimization procedures. XGBoost hyperparameters are optimized via Optuna (30 trials per outer fold) with explicitly stated ranges: n_estimators [100, 500], max_depth [3, 10],

learning_rate [0.01–0.1], min_child_weight [1, 15], gamma [0, 0.3], subsample [0.6–1.0], colsample_bytree [0.6–1.0], reg_alpha and reg_lambda [0, 1], scale_pos_weight [1, 85]. BaggingPU parameters are optimized via exhaustive GridSearch over n_estimators {10, 20, 30, 40, 50} and max_samples {0.1, 0.15, . . . , 0.5} (45 combinations). Nested five-fold CV structure, 50 km spatial block size, random seed (42), and bootstrap configuration (50 models) are specified.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code will be released on GitHub upon acceptance. The geological and geophysical data are compiled from publicly available sources: Geological Survey of Western Australia (GSWA) for 1:500,000 State Interpreted Bedrock Geology and State Geophysical Compilation; Geoscience Australia (GA) for National Gravity Compilation 2019 and Australian Seismological Reference Model [58]. Data access instructions, preprocessing scripts, and complete modeling pipelines will be provided in the code repository.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 comprehensively documents all experimental settings:

   - **Section 3.2**: Dataset statistics (70,200 grid cells, 818 deposits, 1.17% positive rate), 23 predictor features in 5 categories, preprocessing pipeline (missing indicator for $>$5% missingness, Winsorization at 1st/99th percentile, log transform for proximity features, RobustScaler)
   - **Section 3.3**: Variogram analysis (Gaussian model, effective range 31.5 km), 50 km block size (1.59 times the effective range, consistent with 1.25–2.0 times the range in ecological modeling), nested five-fold CV structure, fold distribution ($\sim$14,040 samples, $\sim$164 positives per fold)
   - **Section 3.4**: Evaluation metrics (PR-AUC primary optimization objective, ROC-AUC secondary)
   - **Section 3.5**: Hyperparameter optimization (Optuna 30 trials for XGBoost, GridSearch 45 combinations for BaggingPU)
   - **Section 3.6**: Bootstrap ensemble (50 models, rel_IQR = (Q75-Q25)/mean for uncertainty)
   - **Section 3.7**: Cross-conformal calibration (20 iterations, Platt scaling, FNR control α=0.15)
   - **Section 3.8**: Practical zone classification (5-tier system with lift-based percentiles at 1%, 5%, 10%, 30% rel_IQR threshold conceptually inspired by resource estimation conventions)

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experimental results are reported as mean $\pm$ standard deviation across five outer folds of nested cross-validation. For example: Coverage 96.08% $\pm$ 1.44% (XGBoost), 96.20% $\pm$ 2.17% (BaggingPU); Outer test PR-AUC 0.195 $\pm$ 0.076 (XGBoost), 0.194 $\pm$ 0.080 (BaggingPU); Outer test ROC-AUC 0.911 $\pm$ 0.028 (both models). The variability captured reflects spatial fold variation in the nested CV design. Section 4.2 explicitly acknowledges that with only five outer folds, formal statistical testing of model differences is limited; the reported standard deviations characterize fold-level variability rather than confidence intervals from a large-sample distribution.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 specifies the computational environment: Intel Core i9-14900K CPU (3.20 GHz), 64 GB RAM, and NVIDIA RTX 4090 GPU, along with software versions (Python 3.12, XGBoost 2.0.3, scikit-learn 1.4.0, Optuna 3.5.0, NumPy 1.26.4). Fixed random seed (42) ensures reproducibility. Appendix A provides additional details on hyperparameter configurations per fold.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research uses publicly available geological and geophysical datasets from government agencies (GSWA, GA) under open data licenses. No human subjects or private data are involved. The research aims to improve mineral exploration efficiency, which has positive societal benefits for critical mineral discovery while potentially minimizing environmental impact through targeted exploration rather than broad-area disturbance.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 4.4 and Section 5 discuss practical implications. Positive impacts include: improved exploration efficiency (11-fold concentration factor) potentially reducing environmental footprint through targeted exploration, better resource allocation for critical mineral discovery essential for energy transition, and transparent uncertainty communication enabling risk-informed decision-making. The framework explicitly excludes low-confidence regions (Zone 4) from exploration consideration, preventing wasted resources. Potential indirect environmental impacts of mining activities are acknowledged in Section 4.5 but are beyond the scope of prospectivity mapping methodology.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The mineral prospectivity mapping framework poses no significant risk for misuse. The predictions indicate geological favorability for mineralization based on publicly available geological and geophysical data and cannot be directly used for harmful purposes. The input datasets are already freely available from government agencies under open data policies.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section 3.2 explicitly credits data sources: Geological Survey of Western Australia (GSWA) for 1:500,000 State Interpreted Bedrock Geology and State Geophysical Compilation; Geoscience Australia (GA) for National Gravity Compilation 2019; Australian Seismological Reference Model [58]. These datasets are publicly available under Australian Government open data licenses permitting research use. Software libraries (XGBoost [8], Optuna [62], scikit-learn) are cited with appropriate references and used under their respective open-source licenses.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code repository (to be released upon acceptance) will include: complete preprocessing and modeling pipelines with configuration files, nested CV implementation with spatial block assignment, hyperparameter optimization scripts for both Optuna and GridSearch, bootstrap ensemble and cross-conformal calibration utilities, and practical zone classification module. Documentation includes README with usage instructions, requirements.txt for dependencies, and example notebooks demonstrating the complete workflow from raw data to zone classification.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or human subjects. All data are geological and geophysical measurements from remote sensing and field surveys compiled by government geological surveys, with deposit labels from official mineral occurrence databases maintained by GSWA.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects. The study uses geological and geophysical datasets and mineral occurrence records from government databases, with no human participation or personal data involved.