Comparative Evaluation of Machine Learning Models for Gold Prospectivity Mapping: A Case Study from Labrador, Canada

Yahya Chebab[1], Kongwen (Frank) Zhang[1], Shuyun Xie[2], Jiangtao Liu[3]

[1] School of Computing, University of the Fraser Valley, 33844 King Road, Abbotsford, BC V2S 7M8, Canada - Yahya.Chebab@student.ufv.ca, Frank.Zhang@ufv.ca
[2] State Key Laboratory of Geological Processes and Mineral Resources (GPMR), Faculty of Earth Sciences, China University of Geosciences, Wuhan 430074, China - tinaxie@cug.edu.cn
[3] Wuhan Center, China Geological Survey (Geosciences Innovation Center of Central South China), Guanggu Ave., Wuhan 430205, China - liujt@mail.cgs.gov.cn

Keywords: Gold Prospectivity, Machine Learning, Fuzzy Weights of Evidence, Data-Driven Mineral Exploration, Remote Sensing, Geochemical Analysis.

Abstract

Machine learning methods are increasingly applied to mineral prospectivity mapping. However, systematic comparisons between modern ML techniques and traditional methods such as Fuzzy Weights of Evidence (FWoE) remain limited. In this study, we evaluated four machine learning models (Logistic Regression (LR), Support Vector Machine (SVM), Backpropagation Neural Network (BPNN), and XGBoost) alongside the FWoE method for gold prospectivity mapping in a remote region of Labrador, Canada. We used 1,159 lake sediment samples analyzed for 91 geochemical and physical properties. Rather than binary classification, we developed a four class system based on gold concentrations: Background (no gold), Low, Moderate, and High potential.

Our results showed that XGBoost achieved the highest macro averaged F1 score (0.279), followed by Logistic Regression (0.270). SVM obtained the highest accuracy (0.724) but this reflects its strong performance on background samples rather than its ability to identify mineralized areas. FWoE scored 0.233 and BPNN scored 0.206. All models performed well on background samples but struggled to distinguish between low, moderate, and high gold classes. Feature importance analysis revealed that geochemical elements including copper, arsenic, and molybdenum were most predictive, though physical properties and field observations also contributed. Our findings indicate that XGBoost is the most effective model for multi class gold mapping, but additional data types such as geological maps and geophysical surveys are needed to improve discrimination between different gold grades.

1. Introduction

Recent research has increasingly applied machine learning techniques to predict mineral occurrences. Studies by Zhang et al. (2023) and Wen et al. (2025) demonstrate that ML methods can identify complex patterns in geological data that traditional approaches might miss. At the same time, conventional methods such as fuzzy weights of evidence continue to perform well in certain contexts (Xie et al., 2023; Zhang and Zhou, 2015).

Most existing studies frame mineral prospectivity as a binary classification problem, distinguishing between mineralized and non mineralized areas. However, exploration companies require more detailed information to make drilling decisions. They need to know whether an area has low grade, medium grade, or high grade potential. Multi class classification is considerably more challenging because the geochemical differences between gold grades can be subtle.

In this study, we compare four machine learning models (Logistic Regression, SVM, BPNN, and XGBoost) with the FWoE method using a geochemical dataset from Labrador, Canada. Our objectives are threefold: (1) to determine which model performs best for predicting four gold classes, (2) to identify which geochemical and physical features are most important for these predictions, and (3) to draw practical conclusions for mineral exploration projects.

2. Study Area and Data

2.1 Where We Worked

The study area is located in northern Labrador, Canada, within the Canadian Shield. This region contains ancient Precambrian rocks and hosts known gold deposits, typically occurring in quartz veins associated with elements such as arsenic and antimony. Glacial activity has exposed bedrock in many areas, making lake sediment sampling an effective exploration method.

2.2 The Dataset

Data were obtained from a regional survey conducted by the Geological Survey of Canada in 1986 (Natural Resources Canada, 2023). The survey collected 1,244 lake sediment samples across 12,634 square kilometers, with an average sampling density of approximately one sample per 10 square kilometers. Each sample was analyzed for 91 properties, including metal concentrations (gold, silver, arsenic, copper), physical properties (organic matter content, pH), and field observations (sediment color, stream characteristics).

Gold concentrations were measured using two methods: a standard method (Au) and a more sensitive method (Au_R). We prioritized Au_R measurements when available due to their higher accuracy. After removing 85 samples lacking gold measurements, the final dataset contained 1,159 samples.

2.3 Creating Our Gold Classes

We developed a four class classification system based on gold concentration thresholds derived from quantiles of positive gold values **(Table 1)**.

Class	Category	Gold Concent-ration (ppb)	Number of Samples	Percentage
0	Background	≤ 0	837	72.2%
1	Low	0 to 1.0	141	12.2%
2	Moderate	1.0 to 2.0	112	9.7%
3	High	> 2.0	69	6.0%
Total			1,159	100%

Table 1. Gold class definitions and distribution.

These thresholds have geological meaning. Class 0 represents background concentrations. Class 1 indicates weak anomalies that may originate from distant sources. Class 2 suggests proximity to mineralization. Class 3 indicates strong anomalies characteristic of near source environments.

3. How We Did the Analysis

3.1 Cleaning Up the Data

Data preprocessing involved several steps. Nine columns with 100% missing values (Bi, Cr, Se, Ca_w, Mg_w, Othr_OF, Infill, Alk_w, SO4_w) were removed. For 26 numeric columns with partial missing data, missing values were imputed using the median. For 17 categorical columns, missing values were imputed using the mode.

Geochemical elements including Ag, As, Cu, Fe, Zn, Pb, W, Mo, Hg, Ni, Ba, Cd, Co, Mn, U, and V underwent log transformation. Values ≤ 0 (below detection) were replaced with 0.001 prior to transformation. All numeric features were then standardized to zero mean and unit variance using StandardScaler.

Sixteen categorical variables with fewer than 15 unique values (Province, Sam_Type, Contam, Bnk_type, Wat_colo, Stm_flow, Sed_colo, and others) were encoded using one hot encoding with drop first treatment. No high cardinality categorical columns were present. After preprocessing, the final dataset contained 1,159 samples with 72 features.

3.2 The Machine Learning Models

Four machine learning models were implemented:

Logistic Regression (LR): A linear classifier serving as a baseline model. Multinomial logistic regression was used with the LBFGS solver and maximum iterations of 1000.

Support Vector Machine (SVM): A non linear classifier using the radial basis function kernel. Probability estimates were enabled for comparison with other models.

XGBoost: A gradient boosted tree ensemble with 100 estimators and a learning rate of 0.1. Multi class log loss was used as the evaluation metric.

Backpropagation Neural Network (BPNN): A feedforward neural network with architecture: Dense(64) + Dropout(0.3) + Dense(32) + Dropout(0.3) + Dense(16) + Dropout(0.2) + Dense(4) with softmax activation. Training used the Adam optimizer (learning rate 0.001) for up to 200 epochs with early stopping (patience 20) based on validation loss.

3.3 The Fuzzy Weights of Evidence Model

A custom FWoE model was developed for multi class classification using a One versus Rest strategy. For each class, the model splits each feature into seven fuzzy sets based on percentiles. It then calculates contrast (C) and studentized contrast (tau) for each fuzzy set, followed by applying a sigmoid function to the tau values to obtain membership strengths. Weights of evidence are calculated for the transformed memberships, and finally, evidence from all features is combined to generate class probabilities (Zhang and Zhou, 2015).
Triangular membership functions were used with parameters k=1.5 and shift=0.0, following standard practice in mineral exploration (Zhang and Zhou, 2015).

3.4 How We Tested the Models

Data were split into training (80%, 927 samples) and test (20%, 232 samples) sets using stratified sampling to preserve class proportions. Three fold stratified cross validation was used for model selection and hyperparameter tuning.

Due to class imbalance, accuracy alone is insufficient for model evaluation. A model predicting all samples as background would achieve 72% accuracy while being useless for exploration. Therefore, macro averaged precision, recall, and F1 score were used as primary metrics. Macro averaging calculates metrics for each class separately and averages them, ensuring all classes contribute equally regardless of their frequency.

4. Results

4.1 Confusion Matrix Analysis

Table 2 presents a summary of correct classifications for each model. Figure 1 shows the confusion matrix for XGBoost, which achieved the best overall performance. Full confusion matrices for all five models are provided in Appendix A.

Model	Class 0 Correct	Class 1 Correct	Class 2 Correct	Class 3 Correct
Logistic Regression	153	2	1	1
SVM	168	0	0	0
XGBoost	157	2	1	1
FWoE	141	2	0	1
BPNN	132	0	0	1

Table 2. Summary of correct classifications by class for all models.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-25-2026 | © Author(s) 2026. CC BY 4.0 License.

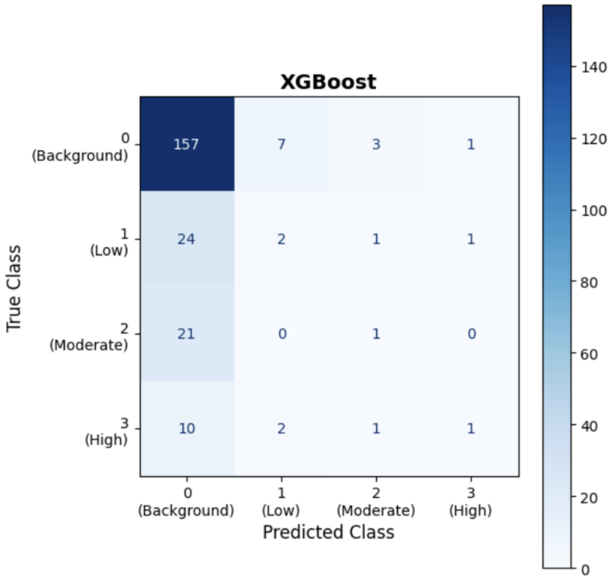

Figure 1. Confusion Matrix for XGBoost

XGBoost correctly classified 157 background samples, 2 low grade samples, 1 moderate grade sample, and 1 high grade sample. While performance on mineralized classes remains low, XGBoost at least attempted to predict these minority classes, unlike SVM and BPNN which failed entirely on Classes 1 and 2.
All models performed well on Class 0 (background), with SVM achieving perfect classification of background samples. However, performance on mineralized classes was poor across all models. SVM and BPNN failed to identify any Class 1 or Class 2 samples, while XGBoost and Logistic Regression showed limited success with only 1-2 correct predictions per mineralized class.

4.2 How the Models Compared

Table 3 presents macro averaged performance metrics for all models on the test set. XGBoost achieved the highest macro F1 score (0.279) and precision (0.356). Logistic Regression followed closely with an F1 score of 0.270. SVM obtained the highest accuracy (0.724) but its macro F1 score was only 0.210, reflecting its strong performance on the majority class at the expense of mineralized classes. FWoE scored 0.233 and BPNN scored 0.206.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.677	0.320	0.275	0.270
SVM	0.724	0.181	0.250	0.210
XGBoost	0.694	0.356	0.281	0.279
FWoE	0.621	0.223	0.246	0.233
BPNN	0.573	0.200	0.214	0.206

Table 3. Test set results for all models (macro averaged).

Cross validation results on training data showed similar patterns. SVM achieved the highest cross validated accuracy (0.725) while XGBoost achieved the highest cross validated macro F1 (0.354). Performance degradation from cross validation to test set was minimal, indicating limited overfitting.

4.3 Breaking Down Performance by Class

Table 4 shows per class F1 scores for all models. All models performed well on Class 0 (background), with F1 scores ranging from 0.765 to 0.840. Performance on mineralized classes was substantially lower. For Class 1 (low grade), the highest F1 score was 0.114 (Logistic Regression). For Class 2 (moderate grade), the highest F1 score was 0.071 (LR and XGBoost). For Class 3 (high grade), the highest F1 score was 0.118 (XGBoost).

Model	Class 0	Class 1	Class 2	Class 3
Logistic Regression	0.814	0.114	0.071	0.080
SVM	0.840	0	0	0
XGBoost	0.826	0.103	0.071	0.118
FWoE	0.786	0.077	0	0.071
BPNN	0.765	0	0	0.061

Table 4. Per-class F1 scores for all models on the test set.

SVM and BPNN failed to predict any Class 1 or Class 2 samples correctly (even Class 3 for SVM), instead assigning all such samples to the background class.

4.4 What Features Mattered Most

We used XGBoost's built in feature importance to see which variables helped the most **(Table 5)**.

Rank	Feature	Importance
1	Cu	0.037440
2	F_w	0.037297
3	Cd	0.036698
4	Sed_colo_Bf Bn	0.032613
5	As	0.032280
6	Mo	0.031346
7	Mn	0.031160
8	Sed_colo_Gy Blu	0.026715
9	LOI	0.024919
10	Pb	0.024795
11	W	0.024732
12	Stm_Phys_Penpln	0.024565
13	Bnk_type_Tal/Scr	0.023973
14	Stm_dpth	0.023952
15	Hg	0.023815
16	Bot_Pcpt_Rd Bn	0.023608
17	Co	0.023158
18	Sed_Cmp	0.023051
19	pH	0.022795
20	Bnk_Pcpt_none	0.022745

Table 5. Top 20 feature importances from XGBoost.

Geochemical elements dominated the list. 12 of the top 20 were metal concentrations. But it's interesting that some physical properties (F_w, LOI, pH, stream depth) and field observations (sediment colors, bank types) also made the list. This tells us that just looking at metal concentrations isn't enough. The environment where the sediment came from matters too.

When we grouped features by category, geochemical variables (36 features) had higher average importance (0.0150) than other features (36 features, average 0.0127). So yes, geochemistry is the most important, but other stuff still contributes.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-25-2026 |

4.5 Spatial Analysis of Predictions

Spatial visualization of predictions revealed important patterns **(Figure 2)**. Samples correctly classified as mineralized tended to cluster in specific areas, consistent with the geological understanding that gold deposits are not randomly distributed but controlled by specific geological settings.

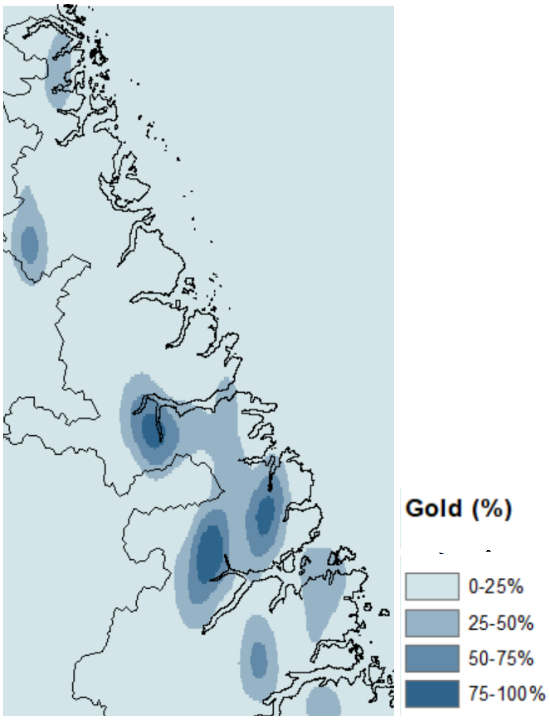

Figure 2. Spatial Distribution of True Gold Labels in Test Set

This map shows where the actual gold bearing samples are located. You can see they cluster in certain areas.

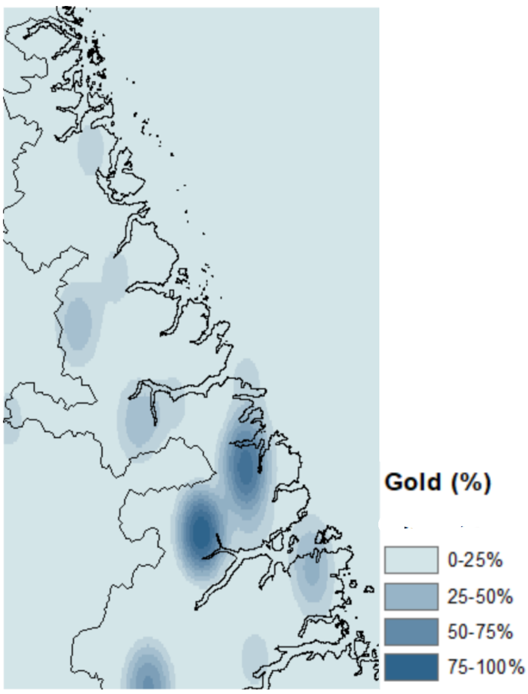

Figure 3. XGBoost Predictions

XGBoost predictions for high grade samples formed coherent clusters that align with areas of known mineralization **(Figure 3)**. The concentration of high grade predictions in specific areas is desirable for exploration targeting, as it provides clear priorities for follow up work. Spatial prediction maps for the other models (Logistic Regression, SVM, FWoE, and BPNN) are provided in Appendix B for comparison.

5. Discussion

5.1 Model Performance Analysis

XGBoost achieved the best overall performance. This result can be attributed to its ability to handle mixed data types (numeric and categorical) and capture complex non linear relationships among variables. In gold systems, elements such as arsenic and antimony commonly co occur, and XGBoost can learn these associations effectively. Logistic Regression, as a linear model, cannot capture such interactions.

The low absolute F1 scores (maximum 0.279) indicate that multi class gold prospectivity mapping is inherently difficult with geochemical data alone. The geochemical differences between low, moderate, and high grade samples are subtle, and natural variability introduces considerable noise.

SVM and BPNN performed poorly on mineralized classes, likely due to their sensitivity to class imbalance. With only 28% of samples in mineralized classes (Classes 1, 2, and 3 combined) and these samples distributed across three classes, the models had insufficient examples to learn distinguishing features. The SVM's RBF kernel may have overfit to the majority class, while BPNN's dropout regularization, though preventing overfitting, may have reduced model capacity excessively.

FWoE achieved a competitive F1 score (0.233) despite being a traditional method. This reflects its conservative design philosophy: it highlights areas with anomalous pathfinder elements even when gold signals are weak. In exploration, false negatives (missed deposits) are more costly than false positives (barren areas requiring follow up), giving FWoE's approach practical value. However, FWoE assumes conditional independence among evidence layers, an assumption frequently violated in geological systems where elements like As and Sb co occur. This violation can inflate posterior probabilities and overstate confidence.

5.2 Implications for Mineral Exploration

These findings have several practical implications for mineral exploration programs:

Model selection should align with exploration objectives. For regional reconnaissance where the priority is identifying all potential areas, FWoE's conservative approach may be preferred despite lower accuracy. For advanced stage targeting where precise boundaries are needed, XGBoost provides the best discrimination between gold grades.

Geochemical data alone are insufficient. The poor performance on mineralized classes demonstrates that multi class prospectivity mapping requires additional data types. Incorporating geological maps, structural information (faults, shear zones), and geophysical surveys (magnetics, radiometrics) would likely improve discrimination, particularly for high grade targets.

Field observations retain value. The appearance of sediment color, bank type, and bottom characteristics in the top features indicates that qualitative field observations capture information about depositional environments not contained in geochemical measurements. These observations should be systematically collected and retained in exploration databases.

5.3 Limitations and Future Work

Several limitations of this study suggest directions for future research:

Class imbalance severely impacted performance. With only 6% of samples in the high grade class, models had insufficient examples to learn distinguishing characteristics. Techniques such as SMOTE (synthetic minority oversampling) or cost sensitive learning (assigning higher penalties to minority class errors) may improve performance. Additionally, because the classes have a natural order (background < low < moderate < high), ordinal regression may be more appropriate than nominal classification.

Validation may be optimistic. Standard cross validation assumes sample independence, but nearby samples are often spatially correlated. This spatial autocorrelation can lead to overestimated performance. Future studies should employ spatial cross validation, where entire regions are held out during training, to obtain more realistic performance estimates.

Additional data types are needed. While comprehensive geochemically, this dataset represents only one data source used in modern exploration. Integration with geological maps, geophysical surveys, and remote sensing data would likely improve model performance substantially.

Probability calibration deserves attention. For exploration decision making, probability estimates are more useful than hard class predictions. A statement such as "this area has a 40% probability of containing high grade gold" provides more actionable information than a binary classification. Techniques such as Platt scaling can improve probability calibration for models like XGBoost.

6. Conclusions

This study compared four machine learning models and the fuzzy weights of evidence method for multi class gold prospectivity mapping using lake sediment geochemistry from Labrador, Canada. The following conclusions can be drawn:

XGBoost performed best overall for distinguishing between gold grades (macro F1 = 0.279). Its ability to capture complex relationships among variables and handle mixed data types makes it suitable for this task.

All models struggled with mineralized classes. Even the best models achieved F1 scores below 0.12 for low and high grade classes, and below 0.08 for moderate grade samples. This indicates that multi class prospectivity mapping is genuinely difficult with geochemical data alone and requires additional geological information.

Geochemistry is most important but not sufficient. Elements including Cu, As, Mo, and Mn were the most predictive features. However, sediment characteristics (color, bank type) also contributed, confirming that field observations provide valuable information.

FWoE remains relevant. Its F1 score of 0.233 demonstrates that traditional methods are not obsolete. For regional exploration where minimizing false negatives is critical, FWoE's conservative approach has continuing value.

Model selection depends on exploration stage. XGBoost is recommended for detailed target definition, FWoE for regional screening, and Logistic Regression when interpretability is paramount.

Mineral exploration remains challenging, and no single model provides a complete solution. The most effective approach likely involves multiple methods, integration of diverse data types, and recognition that models are tools to support geological interpretation rather than replacements for it.

References

Natural Resources Canada. (2023). National Geochemical Reconnaissance Open File 2794: Northern Labrador regional survey. Canada-Newfoundland Mineral Development Program. https://geochem.nrcan.gc.ca/cdogs/content/ngr/ngr_2794_e.htm

Wen, W., Yang, F., Xie, S., Wang, C., Song, Y., Zhang, Y., & Zhou, W. (2025). Quantitative estimation models of the relationship between aeolian sand and soil geochemical composition based on machine learning algorithm, Applied Geochemistry, 191:106528, 10.1016/j.apgeochem.2025.106528. https://www.sciencedirect.com/science/article/abs/pii/S0883292725002513?via%3Dihub

Xie, S., Li, Y., Wang, X., Liu, Z., Ma, K., & Ding, L. (2023). Quantitative prediction of potential areas likely to yield Se-rich and Cd-low rice using fuzzy weights-of-evidence method. Science of The Total Environment, 889, 164015, 10.1016/j.scitotenv.2023.164015. https://www.sciencedirect.com/science/article/abs/pii/S0048969723026360?via%3Dihub

Zhang, Y., Ye, X., Xie, S., Dong, J., Yaisamut, O., Zhou, X., & Zhou, X. (2023) Prediction of Au-Polymetallic Deposits Based on Spatial Multi-Layer Information Fusion by Random Forest Model in the Central Kunlun Area of Xinjiang, China, Minerals, 13, 1302, https://doi.org/10.3390/min13101302.

Zhang, Z., & Zhou, X. (2015). Mineral prospectivity mapping with weights of evidence and fuzzy logic methods. Journal of Intelligent & Fuzzy Systems, 29(6), 2383-2391, 10.3233/IFS-151967. https://journals.sagepub.com/doi/10.3233/IFS-151967

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-25-2026

Appendix A: Full Confusion Matrices

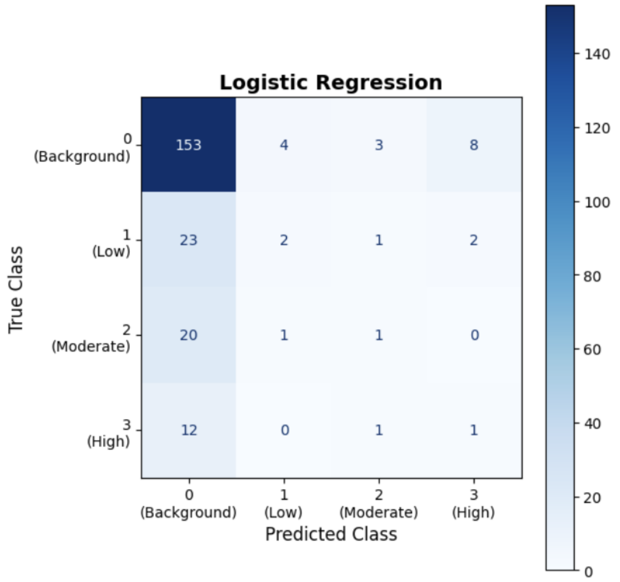

Figure A1. Confusion Matrix for Logistic Regression

Logistic Regression correctly classified 153 background samples but identified only 2 low grade, 1 moderate grade, and 1 high grade sample. Most mineralized samples were misclassified as background, demonstrating the model's difficulty with minority classes despite its interpretability advantages **(Figure A1)**.

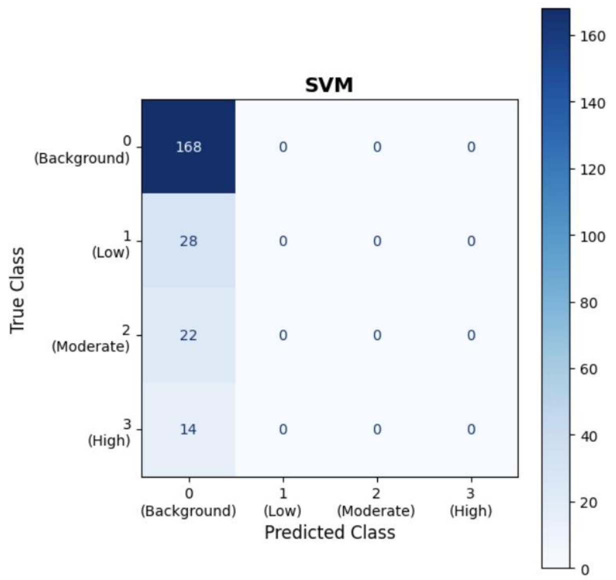

Figure A2. Confusion Matrix for SVM

SVM achieved perfect classification of all 168 background samples but failed to identify any low, moderate, or high grade samples. This confirms that SVM is unsuitable for imbalanced multi class problems, as it simply predicts the majority class for all samples **(Figure A2)**.

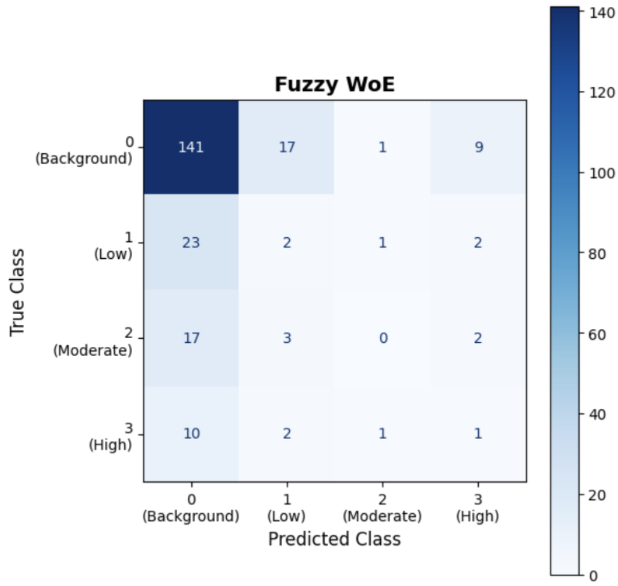

Figure A3. Confusion Matrix for FWoE

FWoE correctly classified 141 background samples, 2 low grade samples, and 1 high grade sample, with no correct moderate grade predictions. Its predictions were more distributed across classes than SVM or BPNN, reflecting its conservative design philosophy **(Figure A3)**.

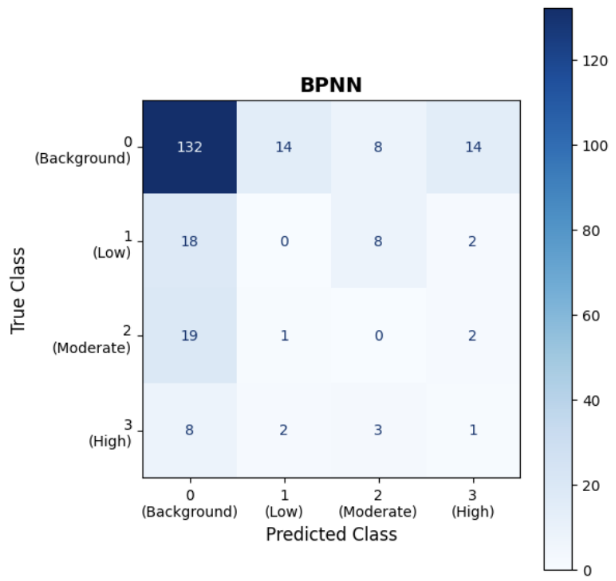

Figure A4. Confusion Matrix for BPNN

BPNN correctly classified 132 background samples (the lowest among all models) and identified only 1 high grade sample, with no correct low or moderate grade predictions. This indicates that the neural network architecture struggled with this dataset, possibly due to limited training data **(Figure A4)**.

Appendix B: Supplementary Spatial Prediction Maps

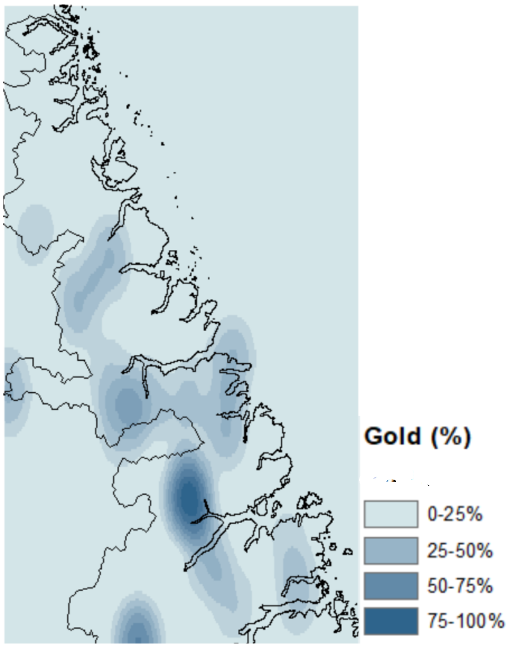

Figure B1. Logistic Regression Predictions

Logistic Regression identified some mineralized areas, but its high grade predictions were spatially dispersed. This lack of clustering would make drill targeting decisions more difficult compared to XGBoost's more concentrated predictions **(Figure B1)**.

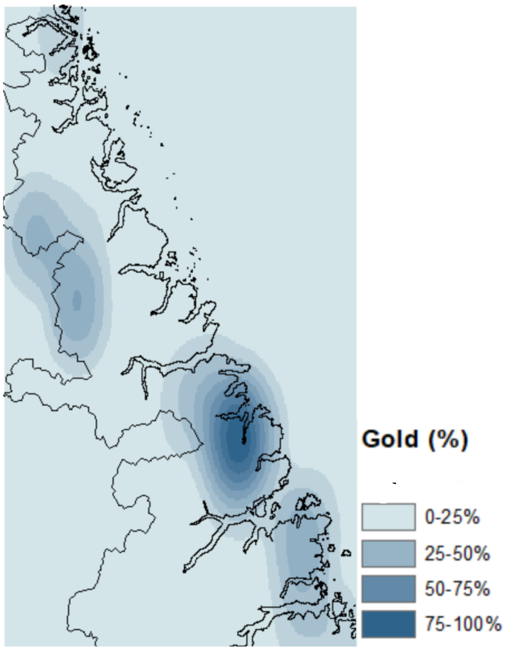

Figure B2. FWoE Predictions

FWoE produced smoother, more distributed predictions characteristic of fuzzy logic methods. While this approach reduces the risk of missing targets, it provides less spatial precision for drill targeting **(Figure B2)**.

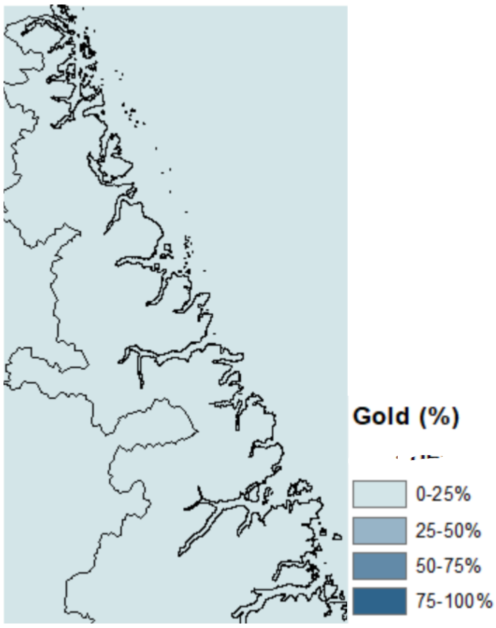

Figure B3. SVM Predictions

SVM predictions were dominated by the background class, with essentially no high grade predictions. This visualization confirms the quantitative results showing SVM's complete failure on mineralized classes **(Figure B3)**.

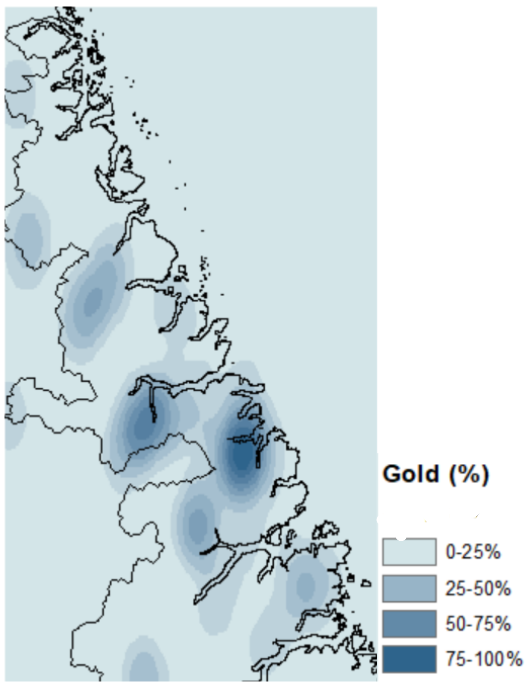

Figure B4. BPNN Predictions

BPNN showed scattered predictions with poor correspondence to actual gold locations, consistent with its low F1 scores on mineralized classes **(Figure B4)**.