HACKATHON 2025

AI for Mineral Targeting

TEAM

anveshan AI

TECHNICAL REPORT

TEAM MEMBERS

Dr. AJOY KUMAR PADHI

BIBHU PRASAD DAS

VENKATESWAR REDDY C

ajoykumarpadhi.amd@gov.in

bibhudas.amd@gov.in

venkateswarreddy.amd@gov.in

TABLE OF CONTENTS

I. Executive Summary

This report presents the development and application of a machine learning–based framework for Rare Earth Element (REE) mineral prospectivity mapping across selected parts of the Dharwar Craton. Integrating domain knowledge with geospatial datasets and ensemble learning methods, the study aims to support more targeted, data-driven mineral exploration strategies in granitic and pegmatite-rich terrains.

Project Scope and Data Curation

The modelling framework was built upon a curated dataset of 309 labeled samples from three key mineralized zones: **Obaganpalli**, **Koppal**, and **Ghattihosahalli**. These zones represent diverse REE enrichment settings within the **Peninsular Gneissic Complex (PGC)** and **Closepet Granite**. Bedrock and C-horizon soil samples from these areas were used to define positive training points, forming the foundation for the prospectivity model.

To construct a robust dataset, two buffer sizes—**1 km and 3 km**—were applied around known mineralized locations, with a **2 km × 2 km grid** used for resampling. Negative labels were generated using two strategies:

- **Soil-informed sampling (SS01)** based on locations with TREE concentrations below 200 ppm.
- **Random spatial sampling (TS01–TS05)** from areas distant from known mineralization.

The 3 km buffer datasets consistently produced superior model performance, highlighting the importance of broader geological context in REE prediction.

Feature Engineering and Input Layers

A total of **21 geospatial features** were extracted following the mineral systems approach. These include:

- **Lithological indicators** and **dolerite density**
- **First- and second-order lineament densities**
- **Geophysical layers** (Reduced-to-Pole magnetic anomalies and Bouguer gravity)
- **Radiometric data** (eK, eU, eTh)
- **Stream sediment geochemistry** (e.g., Zr, Nb, TREE, Ba, CaO, SiO_2)

All features were standardized for compatibility across machine learning models.

To validate structural trends used in the modeling, a **fuzzy K-means clustering algorithm** was applied to magnetic and derivative maps. This helped independently identify subsurface lineaments, which were found to align with known geological structures and provided confidence in the structural features used in the MPM workflow.

Model Development and Ensemble Strategy

The study evaluated four machine learning algorithms:

- **LightGBM**

- **XGBoost**
- **Random Forest**
- **Logistic Regression**

To enhance predictive robustness and reduce reliance on any single algorithm, two ensemble strategies were implemented:

- **Simple averaging** of the three best-performing models (LightGBM, XGBoost, RF)
- **Stacked generalization**, using a gradient boosting meta-learner trained on the predictions of the base models

Each model underwent five-fold stratified cross-validation, resulting in a total of 48 probability maps per ensemble configuration.

Model Performance and Results

- **LightGBM** emerged as the best individual model, achieving an **F1-score of 0.85** and **ROC-AUC of 0.98** on the **SS01_3K** dataset.
- **XGBoost** and **Random Forest** followed closely, with Logistic Regression acting as a baseline model.
- **Soil-informed datasets (SS01)** consistently outperformed random-negative datasets, underscoring the importance of geochemically informed label generation.
- The **stacking ensemble** matched LightGBM in accuracy while offering better generalization and reduced variance across folds.

Target Classification and Uncertainty Analysis

A confidence-based classification system was developed by combining model probability with uncertainty, computed as the **standard deviation across 48 bootstrapped model predictions**.

Target zones were classified as:

- **Priority** (probability ≥ 0.87, uncertainty < 0.07)
- **Potential** (probability ≥ 0.87, uncertainty 0.07–0.10)
- **Exploratory** (probability ≥ 0.87, uncertainty 0.10–0.12)

This hierarchical system allows for risk-tiered exploration planning.

High-Priority REE Zones

The models successfully predicted multiple high-confidence REE prospective zones, particularly around granitic and structurally complex areas. Notable clusters include:

- **Rayadurgam–Kalyanadurg**
- **Gangavathi–Yelburga**
- **Gooty–Peapally**
- **Dharamavaram–Uravakonda**
- **Anantapur–Pamidi**

- **Santebennur–Hosadurga**

These areas show strong geological coherence with REE-bearing pegmatite belts.

Key Predictors and Geological Significance

Feature importance analysis revealed the following:

- **LightGBM** ranked stream sediment geochemical: **Nb, Zr, CaO, Ba,** and **second-order lineament density** as top predictors.
- **XGBoost** highlighted stream sediment geochemical indicators like **TREE(Total REE without Y), Zr, Th, Ba**, and gamma ray spectrometry data of **eK** content

These variables align with known geochemical and structural controls on REE mineralization in granite–gneiss terrains.

Conclusions and Recommendations

This study demonstrates the viability of ensemble machine learning for REE prospectivity mapping in complex geological settings. The following key conclusions were drawn:

- The **LightGBM model trained on soil-informed 3 km buffer data (SS01_3K)** is recommended for operational use due to its high accuracy and geological relevance.
- **Soil-informed negative sampling significantly improves model performance**, emphasizing the importance of domain expertise in data labeling.
- **Broader spatial buffers (3 km)** provide better context for learning mineralization patterns than narrow buffers.
- **Coarse-resolution stream sediment and soil grids**, particularly near Koppala, failed to capture known mineralization. Future exploration should prioritize **higher-resolution, multi-depth geochemical sampling** in these areas.
- Field validation of model-predicted targets and incorporation of new data into iterative model updates are essential for long-term accuracy.

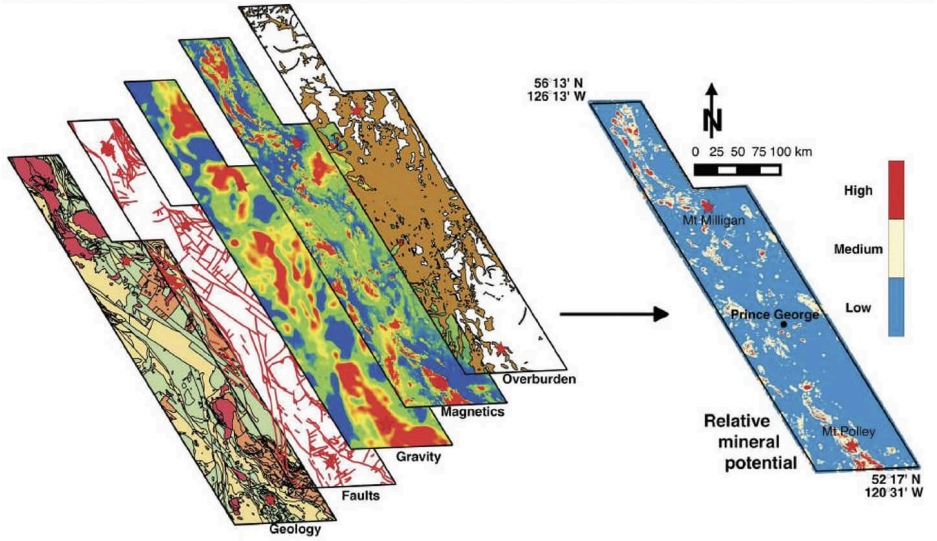
This framework can be scaled to other commodities and regions, offering a replicable blueprint for modern mineral exploration in data-scarce environments.

II. Introduction

Problem Statement

With the global surge in demand for Rare Earth Elements (REEs), particularly for their application in green energy technologies and strategic sectors, there is an urgent need for precise and scalable exploration tools. Traditional exploration techniques often fail to resolve the complex geochemical dispersion patterns and subtle mineralization signatures associated with REEs. This project aims to develop machine learning-based mineral prospectivity mapping (MPM) workflows to delineate REE prospective zones in the Koppal region and adjacent areas within the Western Dharwar Craton.

We adopt an ensemble learning approach that integrates multiple base classifiers—Random Forest, XGBoost, LightGBM, and Logistic Regression—along with a stacking strategy to combine their predictions into a meta-model. The models are trained using integrated geochemical, geological, and geophysical predictor variables. All processing and modeling steps are implemented in Python using libraries such as `rasterio`, `pandas`, `scikit-learn`, `xgboost`, `lightgbm`, and visualized through QGIS.



Schematic of ML based mineral prospectivity mapping

Objectives

1. **To Implement an Ensemble-Based MPM Workflow**: Combine diverse classifiers (RF, XGBoost, LightGBM, Logistic Regression) using stacking and averaging strategies for REE prospectivity mapping.
2. **To Assess Predictive Accuracy Across Sampling Strategies**: Compare performance between datasets with 1 km and 3 km buffers around known mineralization, and negative labels from soil-informed and random background points.
3. **To Incorporate Geological Insight**: Utilize domain-specific geoscientific predictors relevant to REE mineral systems, including pegmatite zones, lithology, and radiometric signatures.
4. **To Generate High-Resolution REE Prospectivity Maps**: Produce interpretable outputs for targeting future exploration efforts.

Methodology

Data Collection:

1.
    - Extract geochemical data from stream sediment and soil-C horizon samples.
    - Compile geological and geophysical layers relevant to REE mineralization (e.g., lithology, Th-U-K radiometric data).
    - Utilize known mineralized locations for generating positive training labels and apply buffered zones to augment sample size.
2. **Data Preparation**:
    - Clean and preprocess datasets using `pandas`, with standard scaling and imputation as required.
    - Ensure CRS uniformity and raster stacking using `rasterio`.
3. **Model Development**:
    - Train base learners (RF, XGBoost, LightGBM, Logistic Regression) with 5-fold cross-validation.
    - Use simple averaging and gradient-boosted stacking for ensemble prediction.
    - Evaluate models using metrics such as ROC-AUC, F1-score, and confusion matrices.
4. **Model Evaluation**:
    - Use 3 km averaged maps for robust comparison.
    - Calibrate the meta-model based on base learner outputs.
    - Visual validation against known REE occurrences.

Map Generation

Prospectivity maps were generated for each model and for the final ensemble outputs using the prediction probabilities. Raster-based outputs were visualized using QGIS to interpret spatial patterns of prospectivity. The stacking meta-model yielded smoother and more geologically consistent maps in comparison to individual classifiers. Zones proximal to known pegmatitic intrusions and shear zones were better highlighted in ensemble maps.

Jupyter Notebooks

The full machine learning workflow has been documented in modular Jupyter notebooks:

- **Data Ingestion & Preprocessing**
- **Feature Selection & Transformation**
- **Model Training & Cross-Validation**
- **Prediction and Raster Generation**
- **Map Visualization with QGIS**

This ensures transparency, reproducibility, and ease of extension for future data integration.

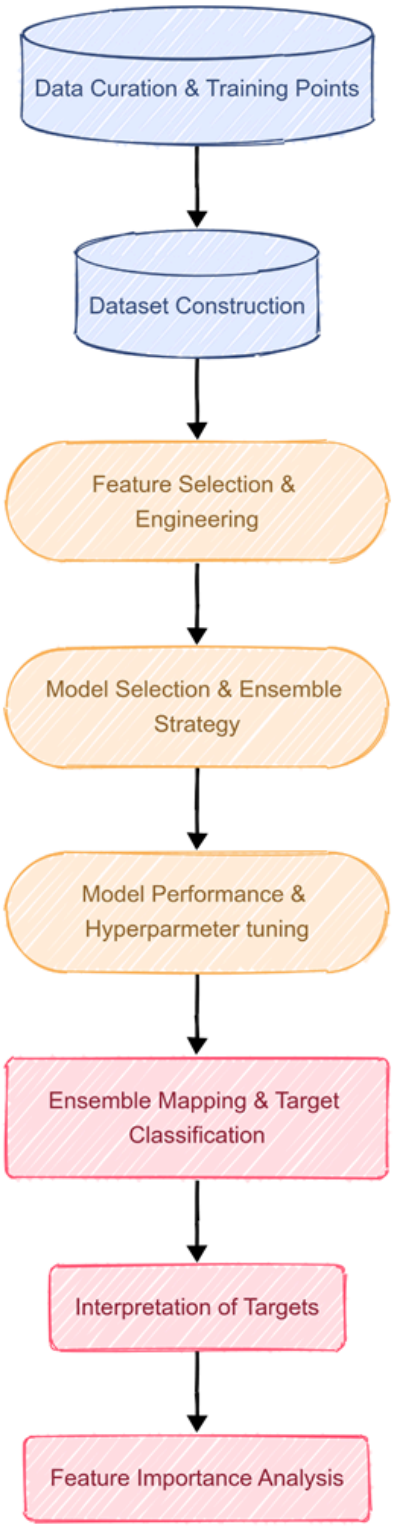
Deliverables

1. **Ensemble Machine Learning Models**: Trained models including RF, XGBoost, LightGBM, and meta-model stacking classifier.

2. **REE Prospectivity Maps**: High-resolution raster outputs showing REE potential zones, including 1 km and 3 km buffered scenarios.
3. **Technical Report**: Documentation of methodology, model diagnostics, and limitations.
4. **Jupyter Notebooks**: Python notebooks covering the full analytical pipeline.
5. **Recommendations for Future Work**: Based on model uncertainty and field validation gaps.

This study demonstrates the application of ensemble machine learning strategies for REE prospectivity mapping in geologically complex terrains. While the models showed strong performance in delineating high-prospectivity zones, limitations exist due to sparse positive labels and coarse geochemical sampling. Ground validation remains crucial, and efforts must focus on acquiring additional labeled data through drilling or trench sampling to reload and retrain models for higher predictive accuracy.

ML PIPELINE


Data Curation & Training Points
Dataset Construction
Feature Selection & Engineering
Model Selection & Ensemble Strategy
Model Performance & Hyperparmeter tuning
Ensemble Mapping & Target Classification
Interpretation of Targets
Feature Importance Analysis

III. Synthesis of Exploration Data for Model Training

In order to develop a reliable machine learning (ML) model for mineral prospectivity mapping, it is essential to curate high-quality training data rooted in real-world mineralization. To this end, an extensive review and synthesis of legacy geological reports and exploration documents was undertaken. While hundreds of reports were examined, only three mineralized locations—**Obaganpalli**, **Koppala**, and **Ghattihosahalli**—provided detailed, structured, and geochemically significant information suitable for model training. These sites span distinct geological domains and host rocks within the Peninsular Gneissic Complex (PGC) and Closepet Granite, offering insights into diverse REE and critical metal enrichment styles. The following section presents a synthesized summary of these locations, highlighting key geological, structural, and mineralogical features relevant for supervised ML applications in mineral exploration.

Breif Description of known REE Occurences:

Obaganpalli

At Obaganpalli, mineralization is found in syenogranite, syenite, quartz syenite, and pegmatite veins within alkali feldspar granite of the Closepet Granite (PGC-II). The geology includes porphyritic granodiorite, migmatitic gneiss, and monzogranite, with multiple intrusive phases such as dolerite dykes trending N310°–330°, N60°–70°, and E–W. REE and related metals (LREE, Zr, Nb, Ta, Co, Li, Ga) are enriched, with REE contents ranging from **0.09% to 0.63%**, and Nb and Zr up to **235 ppm** and **7533 ppm**, respectively. REE minerals include **allanite** and **zircon**, with associated **copper sulphides** (chalcopyrite, covellite, digenite) and **malachite**. Structural controls involve local shears and S-C fabric. Mineralization is attributed to **late-stage magmatic fluids** exploiting structurally weak zones.

Koppala

In Koppala (Budihallu and Dombarhalli), regolith overlying syenite and phoscorite plugs (mafic to ultramafic) hosts significant REE mineralization. The rocks lack quartz and are enriched in MgO, FeO, P_2O_5, CaO, and low in SiO_2. Regolith samples show **TREE up to 4628.72 ppm**, representing 2–5× enrichment over bedrock. Main REE minerals include **allanite-Ce** and its breakdown products (sphene, melanite, joosteite, fergusonite), and associated phases include **magnetite**, **apatite**, **epidote**, and **sulphides**. Structural features include coarse pegmatite veins and NW-SE to E-W trending dolerite/gabbro dykes. The genesis is linked to **volatile-rich mafic melts** during hybridization and **secondary clay-related enrichment**.

Ghattihosahalli

At Ghattihosahalli, mineralization occurs in pegmatite and quartz veins intruding TTG gneiss near the schist belt (PGC-I). Lithium (70–285 ppm), cesium, niobium, tantalum, and tungsten (up to **1500 ppm**) are the main targets. SEM-EDS confirmed **columbite** and **columbo-tantalite**, with additional presence of **spodumene**, **malachite**, and **molybdenum sulphides**. Alteration includes **potassic**, **propylitic**, **phyllic**, and **argillic** zones with silicification and sulphide-bearing quartz veins. The mineralization is attributed to **late-stage hydrothermal activity** crosscutting the TTG gneiss.

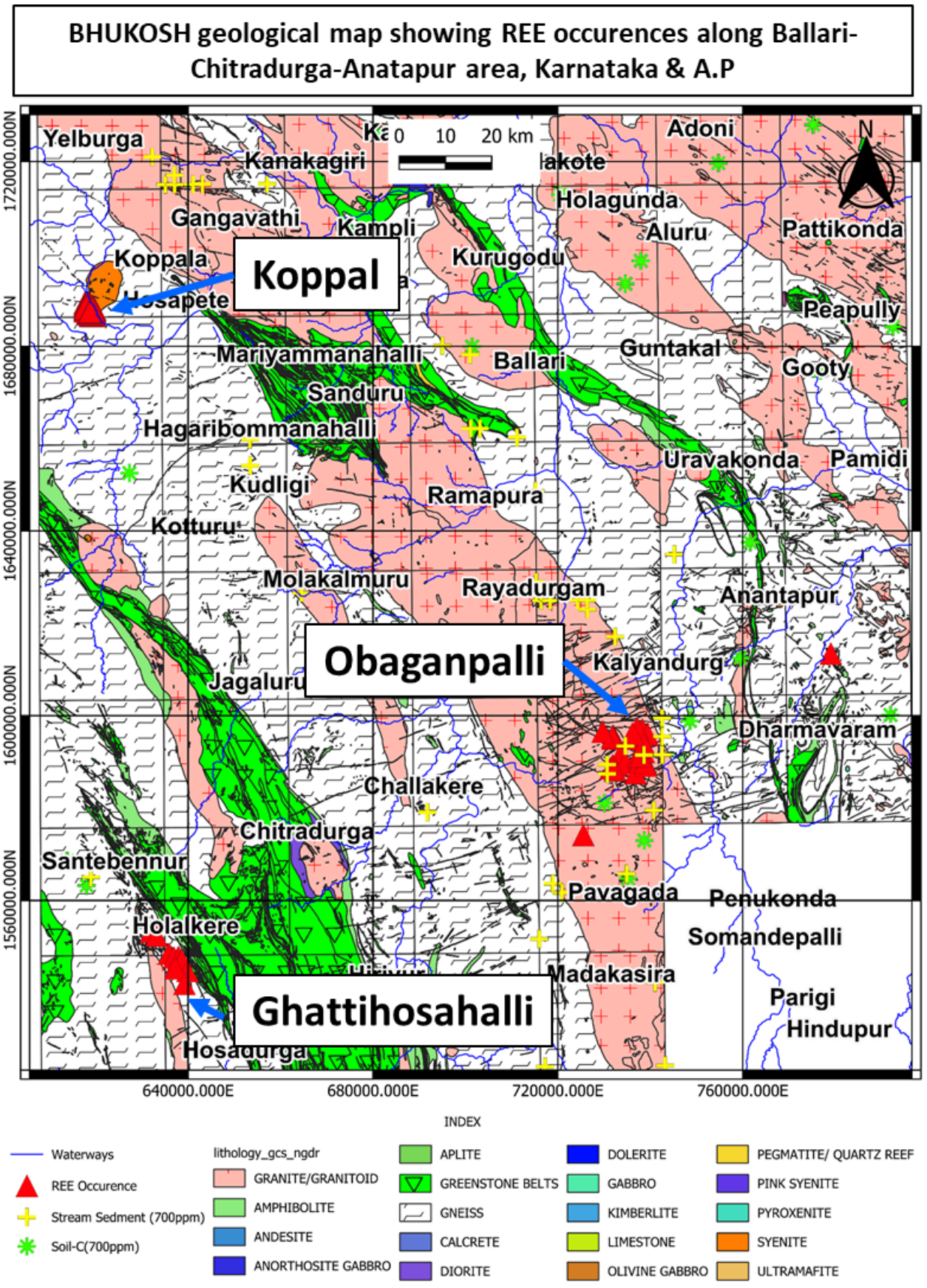
These three localities collectively highlight the significance of structural features, late-stage fluids, and magmatic-hydrothermal evolution in the concentration of REE and critical metals across diverse geological settings.

Data scraped from reports

A total of **259** datapoints were scraped from known mineralization locations. Since those were limited, we also included **50** soil C-horizon samples with **TREE (≥ 500 ppm) ranging from 500ppm to 1500ppm** to bolster the dataset. This a data set of **309 data points were scraped.**

Area	No. of Points	Media	Enriched in	TREE Range (ppm)
Budihallu	83	Bed rock	LREEs (La, Ce), Y	525 – 5153
Dombarhalli	107	Bed rock, Core, Petrochem	LREEs (La, Ce), Y	629 – 5217
Obaganapalli	50	Bed rock (Pegmatite)	LREEs (La, Ce), Y	509 – 1850
Nagalamdike	2	Petrochem	LREEs (La, Ce), Nb	343 – 352
Ghatoshalli	17	Bed rock (Gneiss), Petrochem, Trench	Li,Cs,Nb,Ta,Y	Li : 100 - 285ppm Y upto 1233ppm
Assorted	50	Soil-C Horizon	LREE(without Y)	500-1572

Bhukosh Geological map showing known REE occurrences in the study area

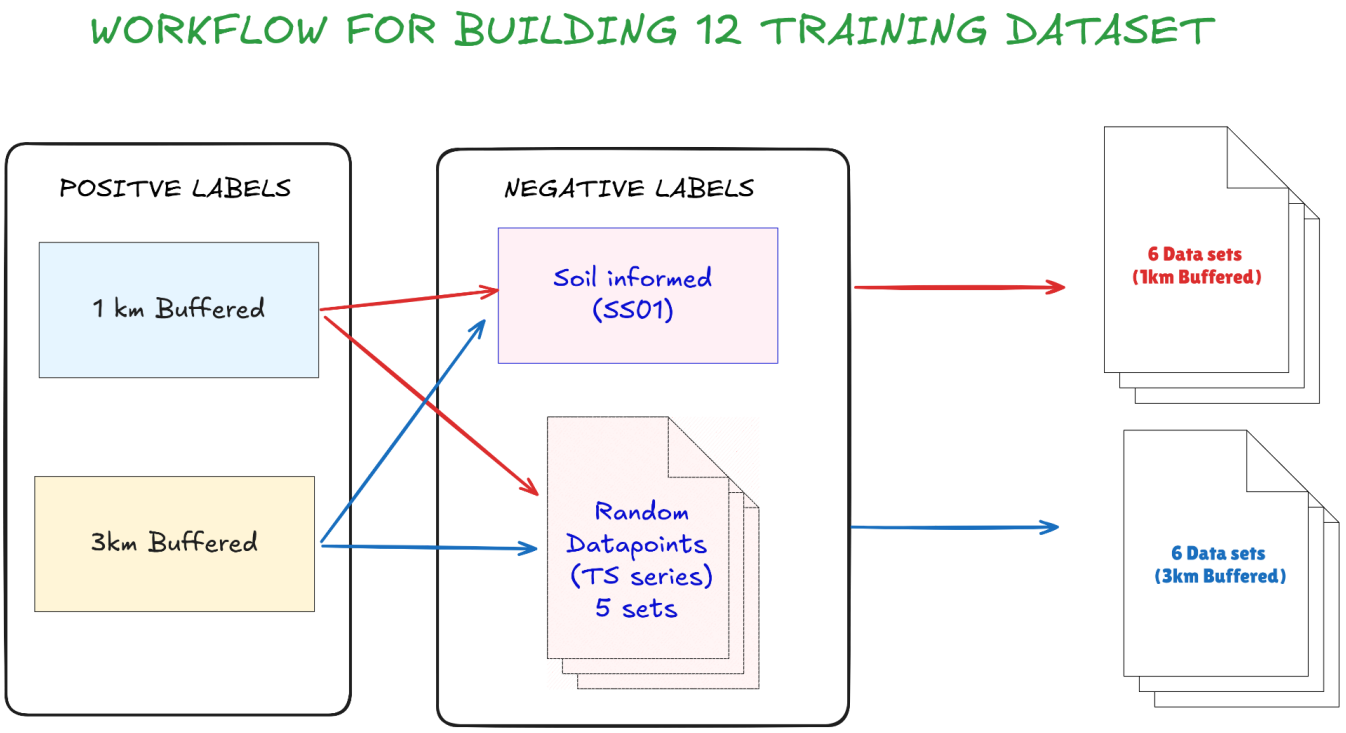


IV. Building the datasets

Dataset Construction Using Multi-Buffer Positive Zones with Soil-Informed & Random Negatives

This chapter describes the methodology followed to build a structured dataset for training and validating machine learning models in rare earth element (REE) mineral prospectivity mapping. The dataset construction was centered on known mineralized locations, buffered at 1 km and 3 km scales, and resampled onto a 2 km × 2 km grid. Two strategies were used to generate negative (non-mineralized) labels: one informed by low-REE soil geochemical data, and another based on spatially random sampling away from known mineralization zones.

The final output includes twelve datasets designed for model training and validation at different spatial buffer scales.



Positive Label Generation

- **Buffering Strategy**: Known mineralized bedrock locations were buffered at:
  - **1 km**, yielding **208** positive grid cells.
  - **3 km**, yielding **530** positive grid cells.
- **Grid Resampling**: Buffered areas were overlaid with a **2 km × 2 km grid**, and intersecting cells were labeled as **positive**.

Negative Label Generation

Due to uncertainty in asserting absolute non-mineralization, two complementary strategies were applied:

1. **Soil-Informed Negative Labels (SS01)**
   Soil sample locations with Total Rare Earth Elements (TREE) < 200 ppm were selected as negative examples.

2. **Random Spatial Negatives (TS01–TS05)**
   Random locations sufficiently distanced from known mineralization zones were selected to avoid contamination by unrecognized mineralization.

Dataset Combinations

For each buffer scenario (1 km and 3 km), six datasets were generated:

- **1 km Buffer Datasets (208 Positives)**:
  - SS01_1k: 208 positive + 200 soil-informed negative labels
  - TS01_1k to TS05_1k: 208 positive + 200 random negative labels (five variations)
- **3 km Buffer Datasets (530 Positives)**:
  - SS01_3k: 530 positive + 500 soil-informed negative labels
  - TS01_3k to TS05_3k: 530 positive + 500 random negative labels (five variations)

Final Dataset Summary

S.No	Dataset Name	Buffer Radius	Positive Labels	Negative Labels	Source of Negatives
1	SS01_1k	1 km	208	200	Soil-informed (TREE < 200 ppm)
2	TS01_1k	1 km	208	200	Random
3	TS02_1k	1 km	208	200	Random
4	TS03_1k	1 km	208	200	Random
5	TS04_1k	1 km	208	200	Random
6	TS05_1k	1 km	208	200	Random
7	SS01_3k	3 km	530	500	Soil-informed with 1km buffer (TREE < 200 ppm)
8	TS01_3k	3 km	530	500	Random
9	TS02_3k	3 km	530	500	Random
10	TS03_3k	3 km	530	500	Random
11	TS04_3k	3 km	530	500	Random
12	TS05_3k	3 km	530	500	Random

This dataset construction framework ensures a balanced and methodical approach to training data preparation. It enables fair comparison between soil-informed and random negative strategies across spatial scales and supports robust model development for REE mineral prospectivity mapping.

V .Feature selection and engineering

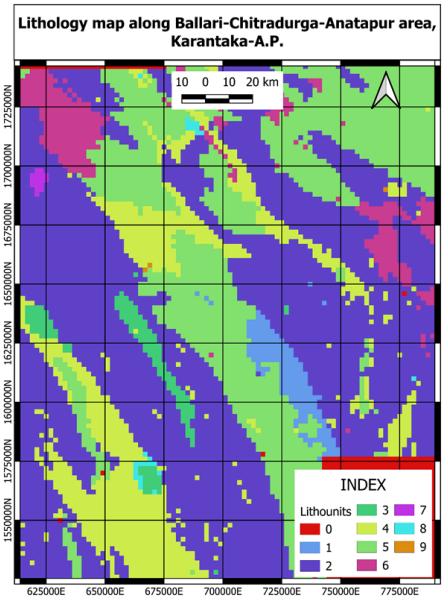
Feature selection and engineering

This chapter describes the derivation and selection of spatial features used in building the feature cube for machine learning-based mineral prospectivity modeling. Features were chosen based on the **mineral systems approach**, incorporating geological understanding from known mineralized occurrences. Special attention was given to distinguishing **granitic- gneiss terrains**, and identifying structural and geochemical controls relevant to REE mineralization.

Al total of **21** features were spatially resampled to a **2 km × 2 km grid** and normalized using **standard normalization** and stacked. **Log normalization** was applied to skewed geochemical variables.

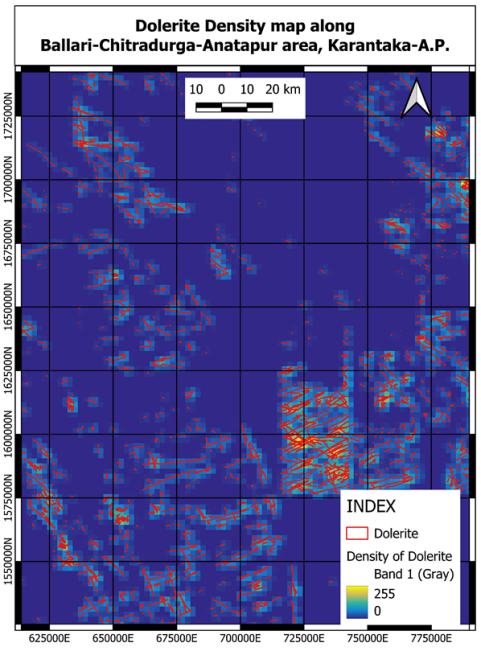
Lithological features

- **Description**: Bhukosh (1: 50000 scale) geological map rasterized from vector geology.
- **Details**: Simplified to **10 lithological classes**, each encoded as an integer value.
- **Purpose**: Distinguish between granitic, gneissic, and associated lithologies relevant to REE host environments.



Dolerite Density features

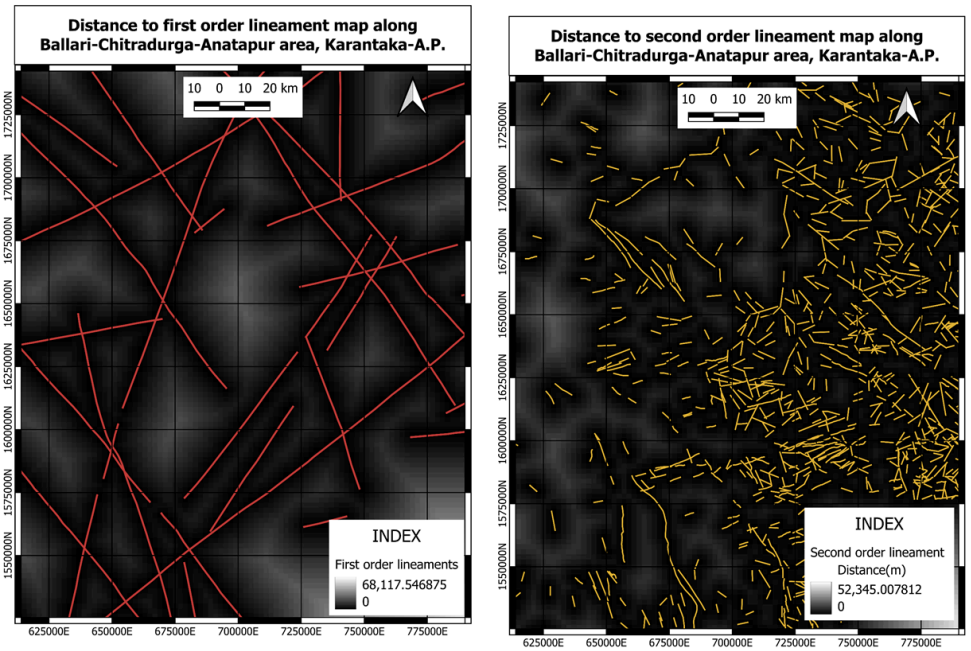
- **Description**: Spatial density of dolerite intrusions.
- **Justification**: Dolerite intrusions are structurally significant and often intersect pegmatites or syenites — both REE-favorable environments.
- **Note**: Represents zones of late-stage structural weakness.

- 

Lineament features

Lineament data were classifed into two: long range **first order lineaments** and short range **second order lineaments**

- **Description**: Lineament data classified by order (regional vs local structures).
- **Processing**: Buffered and rasterized to match the 2 km grid.
- **Relevance**: Potential structural pathways or traps for REE mineralizing fluids.

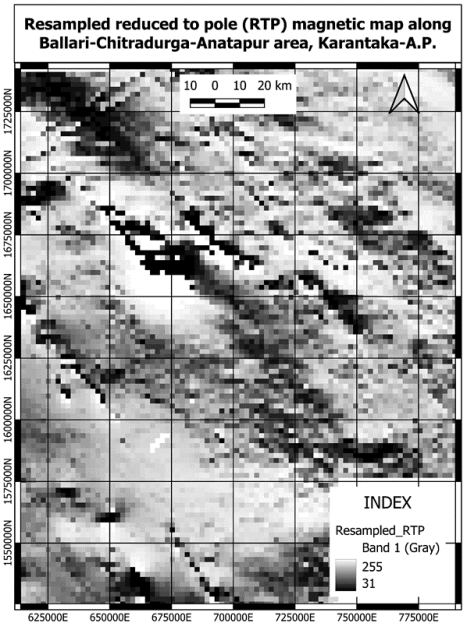
- 

Aeromagnetic features

Aeromagnetic data was processed and was Reduced to Pole

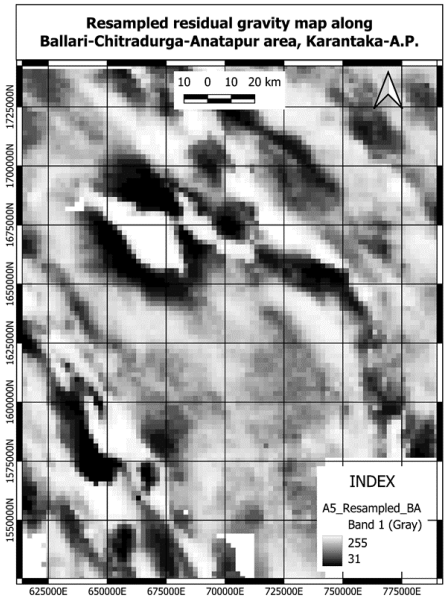
- **Description**: Reduced-to-Pole (RTP) magnetic anomaly data.
- **Reasoning**: Helps delineate magnetic granites and mafic bodies; identifies subsurface structural fabrics.

- **Preprocessing**: RTP derived from processed aeromagnetic data, then resampled to 2 km.



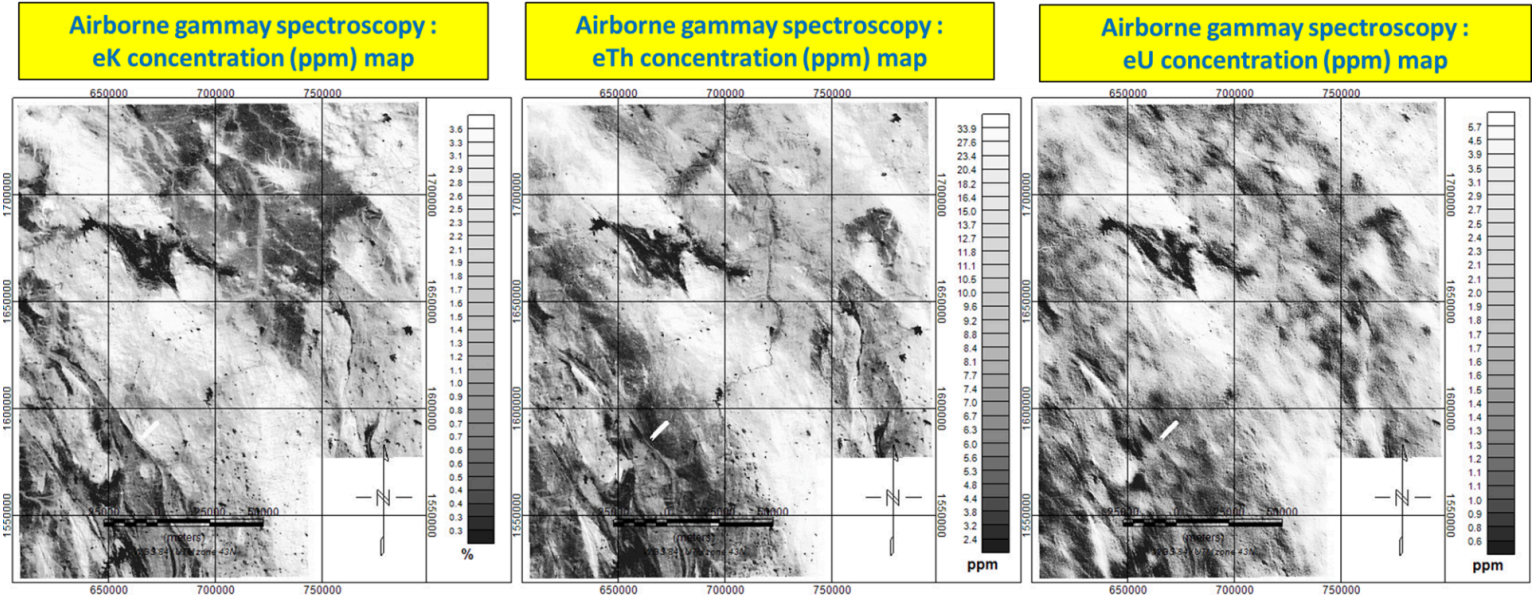
Ground Gravity features

- **Description**: Bouguer Anomaly (residual gravity).
- **Purpose**: Identifies dense vs. less dense zones — critical in differentiating felsic vs mafic intrusives.



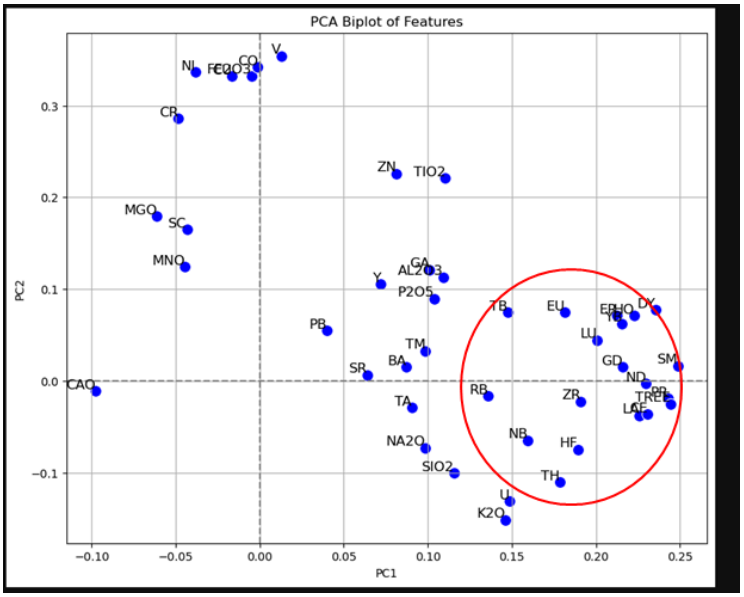
Airborne Gamma-ray spectrometry features

- **Description**: Airborne gamma-ray spectrometry (eK, eU, eTh).
- **Importance**: Crucial for characterizing granitic terrains and mapping K-rich or U-rich granites.
- **Processing**: Log-transformed due to skewness; rasterized and resampled to the modeling grid.

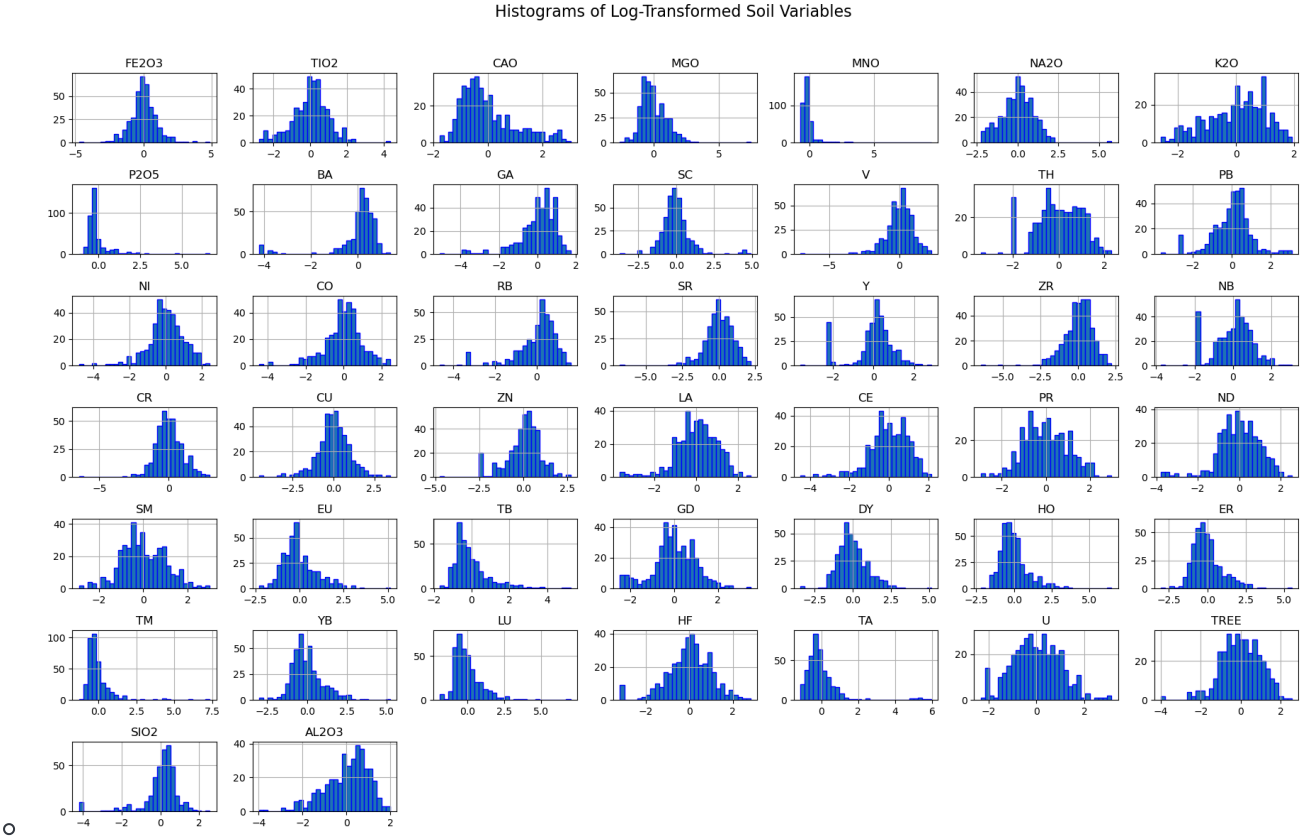


Stream Sediment geochemisty features

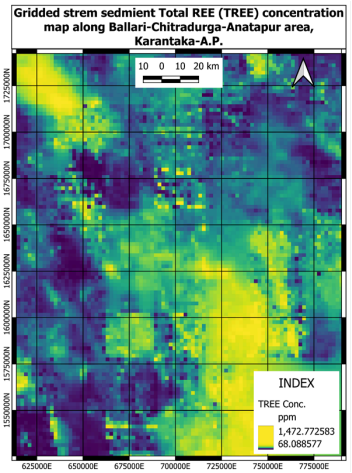
- **Feature selection:** Exploratory data analysis was carried out on stream and soil-c horizon geochemical data to find out related elements to REE; **Principal component analysis** and **correlation matrix** were used to screen elements.

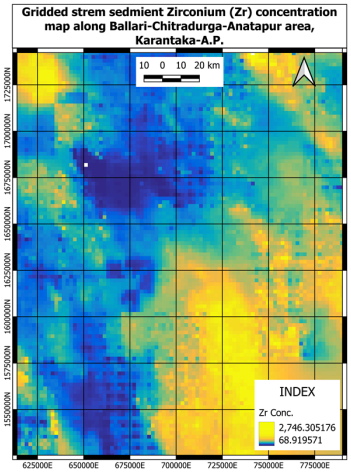


- **Description**: Stream sediment geochemistry — trace and major elements.
- **Sampling Resolution**: Interpolated and rasterized from field data.
- **Selected Stream sediment geochemical elements**:
  - **Zr, Nb, Ce, Th, U, TREE**: Direct REE pathfinders.
  - **Ba, Rb, K2O**: Indicators of feldspar/Mica-content granite differentiation.
  - **MgO, CaO**: Help distinguish mafic/felsic/carbonatite associations.
  - **SiO_2**: Degree of silicity — differentiates more evolved granite types.



- **Preprocessing**: Log normalization and rasterization followed by standard normalization.





Cube Standardization

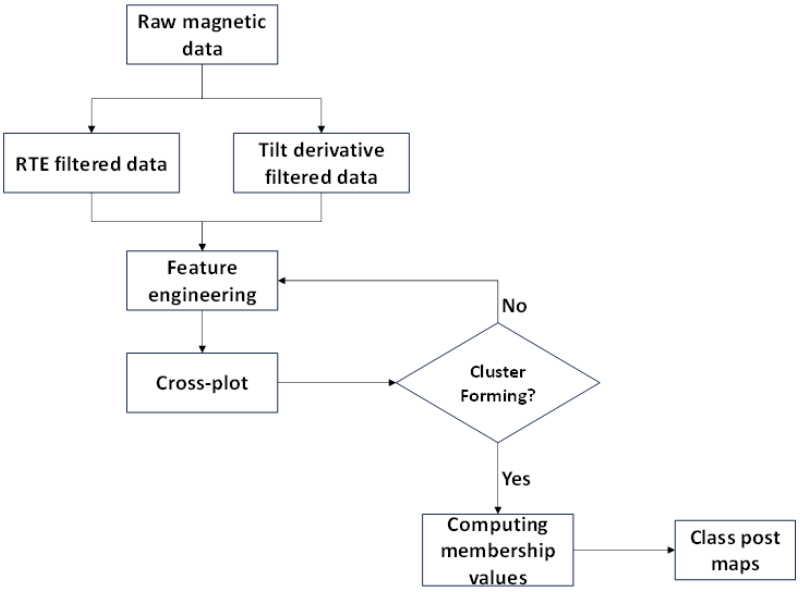
- **Normalization**: All features were standardized (mean = 0, std = 1) prior to model ingestion to ensure compatibility and unbiased model learning.

This structured approach to feature engineering integrates geological, geophysical, and geochemical proxies into a coherent spatial cube. The chosen features aim to capture the **lithological**, **structural**, and **geochemical context** that governs REE mineralization in granite-gneiss terranes. These inputs are essential to ensure that the machine learning models are both **geologically informed** and **statistically robust**.

**

VI. Fuzzy K-means clustering for subsurface lineament mapping

Fuzzy K-means clustering algorithm is unsupervised machine learning algorithm. The algorithm was applied to the magnetic and its derivative maps to mark the subsurface lineaments. The process is depicted by the below flow chart



Flow chart depicting the steps for K-means clustering algorithm.

The raw and/or processed geophysical data have an **intrinsic property of overlapping anomalies**. Here we have chosen **magnetic data** since they are best suited structural analysis. In this approach the problem was addressed as a **classification problem**, to identify and separate geophysical anomalies belonging to in place of the same class. Each data point belongs to **more than one cluster** but with a varying **degree of membership**. In this exercise we have chosen the **highest membership value** for each data point for each cluster. The **Reduced to magnetic equator** and its derivative maps showed **a good cross-correlation**. The data were sampled in a **100m grid** (100m station spacing and 100m line spacing). From the class separation and fuzzy membership assignments a **class post map** was prepared. Each cluster belonging to the same characteristic class have same properties irrespective of the geophysical response at that data point. From the class separation map, we have identified subsurface lineaments and corroborated the results with the available geological data.

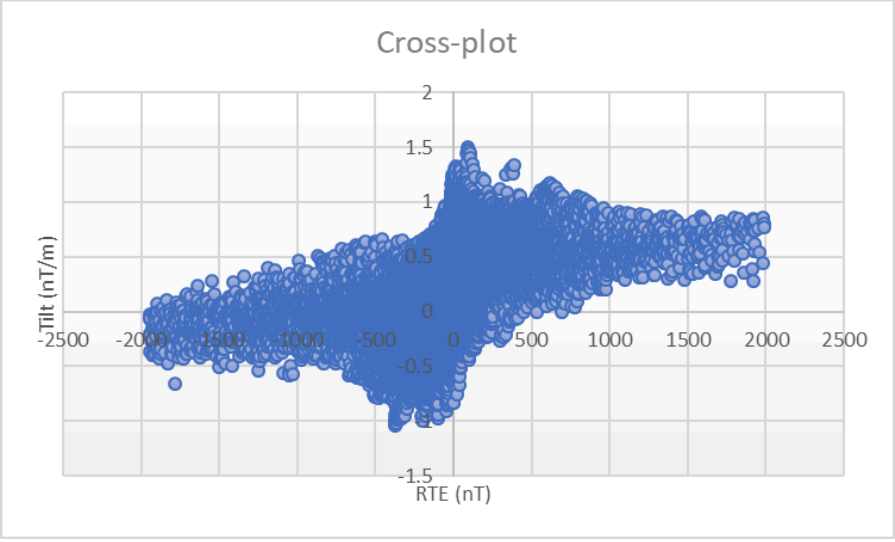


Figure: Cross-plot of RTE and Tilt derivative magnetic data.

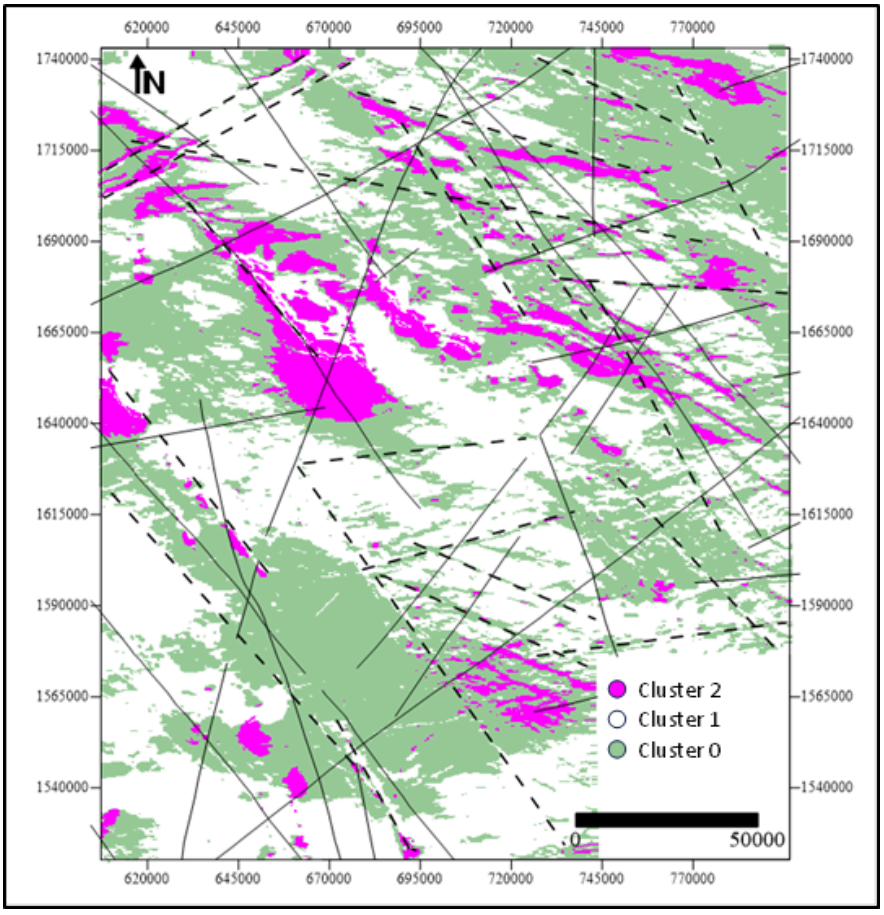


Figure: Class post map of the study area with associated lineaments.

(Dotted lines: inferred from cluster analysis, Solid lines: Marked from NGDR data).

VII. Model selection

Model Selection & Ensemble Strategy

Mineral Prospectivity Mapping (MPM) integrates geological, geochemical, and geophysical datasets to estimate the likelihood of undiscovered mineral occurrences. Machine learning (ML) amplifies this capability by synthesizing diverse predictors — including radiometric anomalies at 1 km and 3 km buffers, soil-informed negative samples, and randomly sampled background sites — to uncover subtle spatial patterns.

Preliminary analysis showed that **3 km buffer models outperformed their 1 km counterparts** in terms of spatial coherence and predictive performance. Building on this insight, **average models** were developed by aggregating the predictions from three strong performers — **Random Forest (RF)**, **LightGBM (LGBM)**, and **XGBoost (XGB)** — trained on the 3 km buffer data. This averaging strategy served as a robust baseline ensemble.

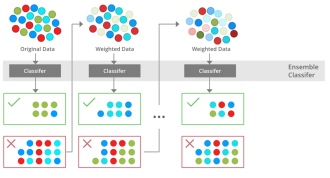
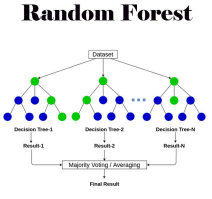
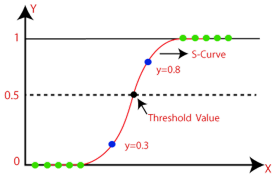
Additionally, to leverage the strengths of all models in a coordinated manner, a **stacked ensemble using Gradient Boosting as the meta-learner** was trained on the **SS01_3K** dataset. This meta-model combined outputs from four diverse base learners — **LightGBM**, **XGBoost**, **Random Forest**, and **Logistic Regression** — into a unified predictive framework. The emphasis was on **combining model intelligence**, not simply choosing the best single model.

Why Avoiding Single-Model Dependency Matters

1. **Workflow-Induced Uncertainty**
   Noise and inconsistencies from data preparation (e.g., interpolation artifacts, spatial resampling, feature engineering) may mislead individual models.
2. **Algorithmic Blind Spots**
   Each model brings its own assumptions — linearity, decision tree binarization, or additive error minimization — which may fail in complex geological environments.
3. **Instability Across Data Splits**
   Dependence on a single train-validation split may yield unreliable generalization.
4. **Limited Pattern Recognition Scope**
   Single models may not capture both local and global patterns. Ensembles improve pattern breadth and depth.

Overview of the Base Models

Block diagram of Base models :

LightGBM	XGBoost	Random Forest	Logistic Regression
			

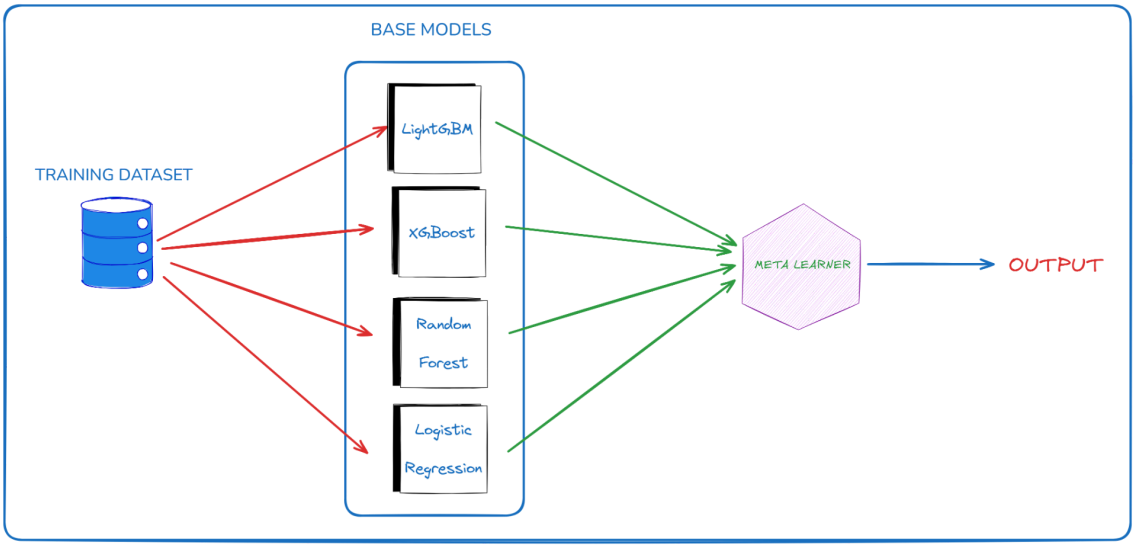
LightGBM	XGBoost	Random Forest	Logistic Regression

Model	Learning Mechanism	Strengths	Limitations
LightGBM	Leaf-wise histogram-based gradient boosting	Fast, low memory, handles large data well	Needs careful tuning, overfit risk
XGBoost	Gradient boosting with regularization	Excellent generalization, regularization aware	Complex, slower on very large data
Random Forest	Ensemble of bootstrapped decision trees	Robust to noise, low variance	Slower in large-scale scoring, class imbalance sensitive
Logistic Regression	Linear model via log-odds optimization	Fast, interpretable, calibrated outputs	Limited to linear relationships

Why Use Gradient Boosting Stacking

- **Bias–Variance Trade-off**: Blending high-variance models (e.g., RF, XGB) with high-bias models (e.g., Logistic Regression) achieves a more balanced generalization.
- **Complementary Strengths**: Tree-based models excel at nonlinearity; linear models capture broad trends.
- **Resilience to Workflow Errors**: Meta-learner smooths over fragilities of individual models.

Stacking Strategy Using Gradient Boosting



Workflow

1. **Base Model Training**
   Twelve base models were trained: four models (LGBM, XGB, RF, LR) on each of the four dataset variants — **TS01_1K**, **TS01_3K**, **SS01_1K**, and **SS01_3K**.
2. **Meta-Feature Construction**
   Out-of-fold prediction probabilities from base models were used to create meta-feature matrices using the **SS01_3K** data.
3. **Gradient Boosting Meta-Learner Training**
   A meta-classifier based on Gradient Boosting was trained to learn combination weights.

4. **Prediction Pipeline**
   New data go through each base model to generate meta-features, which are then input into the meta-learner to generate the final prospectivity score.

Key Benefits of the Ensemble Approach

- **Error Balancing**: Avoids over-reliance on any single model's bias or overfitting tendency.
- **Calibrated Outputs**: Boosted ensembles provide more confident and calibrated predictions.
- **Extensibility**: New models or data types can be added to the stack.
- **Consistency**: Better performance across different spatial regions and data variations.

Summary

- A **Gradient Boosting Stacking Classifier** was trained using **SS01_3K** data.
- **Twelve base models** were trained on 1 km and 3 km buffers from **TS01** and **SS01** datasets.
- **Averaged models** from **RF**, **LGBM**, and **XGB** were created using 3 km data.
- The strategy prioritized **ensemble-based reasoning over single-model selection**, improving robustness, accuracy, and reliability for REE prospectivity mapping.

VIII. Model Performance Analysis

This report analyzes the performance of four machine learning models (LightGBM, XGBoost, Random Forest, and Logistic Regression) for Rare Earth Element (REE) mineral prospectivity mapping across different buffer zones and datasets. All the models hyperparameters were tuned manually or using **Randomized grid search method** to achieve best possible generalization cross checked with learning curves. The analysis covers both **1km** and **3km** buffer zones around mineralized locations using various negative sampling strategies.

Dataset Overview

Buffer Zones:

- **1km buffer**: Models trained with 1km buffer around mineralized locations
- **3km buffer**: Models trained with 3km buffer around mineralized locations

Negative Sampling Strategies:

- **SS01**: Soil-informed negative labels (domain knowledge-based)
- **TS01-TS05**: Random negatives sampled at varying distances from mineralized locations

Model Performance Analysis

1. LightGBM Model Performance

Best Performing Model: LightGBM demonstrates superior performance across all metrics and buffer configurations.

Key Strengths:

- Highest test accuracy across both buffer configurations
- Excellent balance between precision and recall
- Minimal overfitting with reasonable train-validation gaps
- Consistent performance across different negative sampling strategies

1km Buffer Results:

- **Validation Accuracy**: 0.67-0.78 (mean), with SS01_1K achieving highest accuracy (0.78)
- **Test Performance**: SS01_1K dataset shows exceptional performance (0.82 across all metrics)
- **Average TS Performance**: Consistent ~0.71 across test metrics

LightGBM model training and validation scores for 1km buffer datasets

Metric	SS01_1K (mean ± std)	TS01_1K (mean ± std)	TS02_1K (mean ± std)	TS03_1K (mean ± std)	TS04_1K (mean ± std)	TS05_1K (mean ± std)
Valid_accuracy	**0.78** ± 0.03	0.70 ± 0.03	0.71 ± 0.07	0.70 ± 0.06	0.67 ± 0.04	0.70 ± 0.04
train_accuracy	**0.83** ± 0.01	0.77 ± 0.02	0.77 ± 0.02	0.75 ± 0.01	0.79 ± 0.03	0.76 ± 0.02
Valid_precision	**0.81** ± 0.04	0.71 ± 0.06	0.75 ± 0.08	0.72 ± 0.10	0.70 ± 0.06	0.71 ± 0.06
train_precision	**0.87** ± 0.01	0.78 ± 0.02	0.83 ± 0.02	0.78 ± 0.01	0.81 ± 0.02	0.76 ± 0.02
Valid_recall	**0.74** ± 0.03	0.70 ± 0.07	0.63 ± 0.08	0.68 ± 0.08	0.64 ± 0.04	0.72 ± 0.07
train_recall	**0.78** ± 0.03	0.75 ± 0.02	0.69 ± 0.05	0.72 ± 0.03	0.76 ± 0.05	0.76 ± 0.03
Valid_f1	**0.77** ± 0.03	0.70 ± 0.03	0.69 ± 0.08	0.70 ± 0.06	0.67 ± 0.03	0.71 ± 0.03
train_f1	**0.82** ± 0.01	0.76 ± 0.02	0.75 ± 0.03	0.75 ± 0.01	0.78 ± 0.03	0.76 ± 0.02

LightGBM Model Testing Scores for 1km Buffer Datasets

Metric	SS01_1K	TS01_1K	TS02_1K	TS03_1K	TS04_1K	TS05_1K
Test_accuracy	**0.82**	0.73	0.72	0.71	0.68	0.70
Test_precision	**0.82**	0.73	0.73	0.71	0.68	0.71
Test_recall	**0.82**	0.73	0.72	0.71	0.68	0.69
Test_f1	**0.82**	0.73	0.72	0.71	0.68	0.69

3km Buffer Results:

- **Validation Accuracy**: 0.71-0.81 (mean), showing improved performance with larger buffers
- **Test Performance**: SS01_3K achieves outstanding 0.85 across all metrics
- **Average TS Performance**: Improved to 0.75 compared to 1km buffer

LightGBM model training and validation scores for 3km buffer datasets

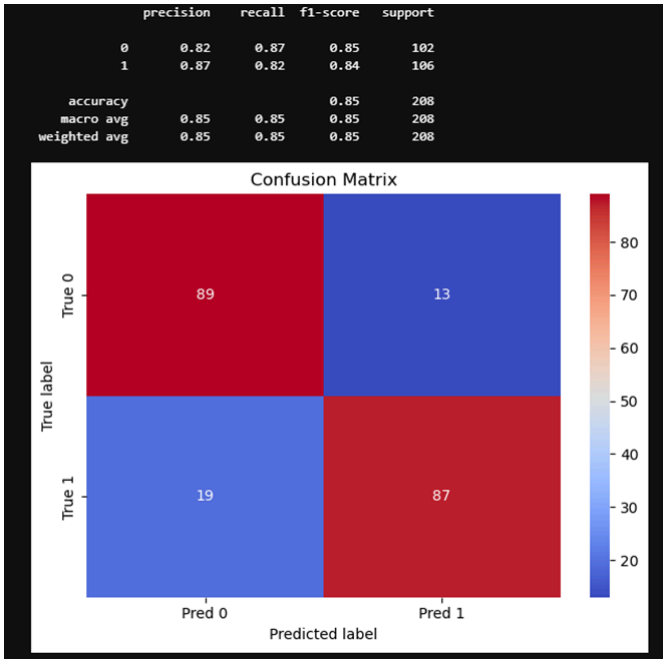
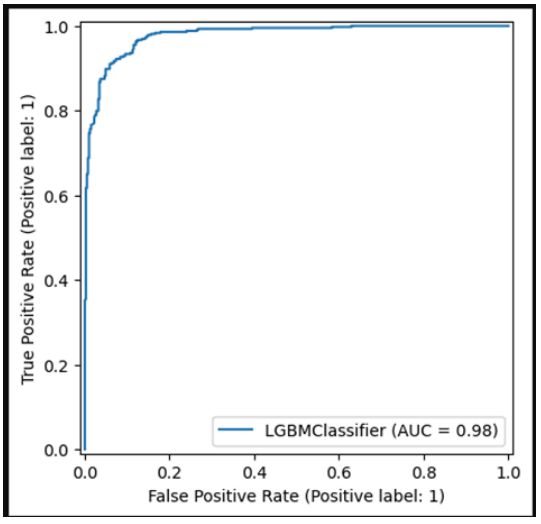
Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
Valid_accuracy	**0.81 ± 0.04**	0.75 ± 0.04	0.74 ± 0.03	0.74 ± 0.04	0.71 ± 0.06	0.72 ± 0.03
train_accuracy	**0.89 ± 0.01**	0.83 ± 0.01	0.81 ± 0.01	0.80 ± 0.01	0.78 ± 0.01	0.78 ± 0.01
Valid_precision	**0.84 ±**	0.76 ±	0.75 ±	0.76 ±	0.73 ±	0.74 ±

Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
	0.03	0.04	0.04	0.05	0.06	0.02
train_precision	**0.92 ± 0.01**	0.84 ± 0.01	0.82 ± 0.01	0.83 ± 0.01	0.80 ± 0.02	0.80 ± 0.01
Valid_recall	**0.77 ± 0.06**	0.76 ± 0.06	0.73 ± 0.05	0.73 ± 0.06	0.69 ± 0.06	0.70 ± 0.07
train_recall	**0.87 ± 0.02**	0.82 ± 0.01	0.80 ± 0.01	0.78 ± 0.01	0.75 ± 0.01	0.76 ± 0.02
Valid_f1	**0.80 ± 0.04**	0.76 ± 0.04	0.74 ± 0.03	0.74 ± 0.04	0.71 ± 0.06	0.72 ± 0.04
train_f1	**0.89 ± 0.01**	0.83 ± 0.01	0.81 ± 0.01	0.80 ± 0.01	0.77 ± 0.01	0.78 ± 0.01

LightGBM Model Testing Scores for 3km Buffer Datasets

Metric	SS01_3K	TS01_3K	TS02_3K	TS03_3K	TS04_3K	TS05_3K
Test_accuracy	**0.85**	0.75	0.75	0.77	0.74	0.75
Test_precision	**0.85**	0.76	0.75	0.77	0.74	0.75
Test_recall	**0.85**	0.75	0.75	0.77	0.74	0.75
Test_f1	**0.85**	0.75	0.75	0.77	0.74	0.75

LightGBM for 3km buffered soil informed dataset

Confusion matrix	ROC-ACUC
	

2. XGBoost Model Performance

Second Best Overall: XGBoost shows competitive performance with good generalization capabilities.

Key Strengths:

- Good performance across different buffer sizes
- Reasonable train-validation gap indicating controlled overfitting
- Strong performance on 3km buffer datasets

1km Buffer Results:

- **Validation Accuracy**: 0.69-0.77 (mean)
- **Test Performance**: Variable across datasets (0.67-0.78), with TS01_1K performing best
- **Average TS Performance**: 0.71 (consistent with other models)

XGBoost Model Training and Validation Scores — 1km Buffer Datasets

Metric	SS01_1K (mean ± std)	TS01_1K (mean ± std)	TS02_1K (mean ± std)	TS03_1K (mean ± std)	TS04_1K (mean ± std)	TS05_1K (mean ± std)
Valid_accuracy	**0.77 ± 0.05**	0.70 ± 0.04	0.70 ± 0.07	0.70 ± 0.05	0.69 ± 0.02	0.70 ± 0.04
Train_accuracy	**0.79 ± 0.01**	0.73 ± 0.01	0.74 ± 0.01	0.74 ± 0.02	0.72 ± 0.01	0.73 ± 0.01
Valid_precision	**0.80 ± 0.06**	0.71 ± 0.04	0.76 ± 0.11	0.72 ± 0.06	0.70 ± 0.02	0.69 ± 0.04
Train_precision	**0.81 ± 0.01**	0.74 ± 0.01	0.79 ± 0.03	0.75 ± 0.03	0.73 ± 0.02	0.73 ± 0.02
Valid_recall	**0.73 ± 0.06**	0.70 ± 0.09	0.61 ± 0.08	0.70 ± 0.08	0.69 ± 0.07	0.73 ± 0.08
Train_recall	**0.76 ± 0.03**	0.72 ± 0.01	0.65 ± 0.03	0.72 ± 0.03	0.71 ± 0.04	0.75 ± 0.02
Valid_f1	**0.76 ± 0.05**	0.70 ± 0.05	0.67 ± 0.07	0.70 ± 0.05	0.69 ± 0.04	0.71 ± 0.05
Train_f1	**0.78 ± 0.02**	0.73 ± 0.01	0.72 ± 0.02	0.74 ± 0.02	0.72 ± 0.01	0.74 ± 0.01

XGBoost Model Testing Scores — 1km Buffer Datasets

Metric	SS01_1K	TS01_1K	TS02_1K	TS03_1K	TS04_1K	TS05_1K
Test_accuracy	**0.78**	0.76	0.67	0.70	0.70	0.72
Test_precision	**0.78**	0.76	0.68	0.70	0.70	0.72
Test_recall	**0.78**	0.76	0.67	0.70	0.70	0.72
Test_f1	**0.78**	0.76	0.67	0.70	0.70	0.72

3km Buffer Results:

- **Validation Accuracy**: 0.72-0.80 (mean), showing improvement with larger buffers
- **Test Performance**: More consistent performance (0.73-0.83)
- **Average TS Performance**: Improved to 0.77

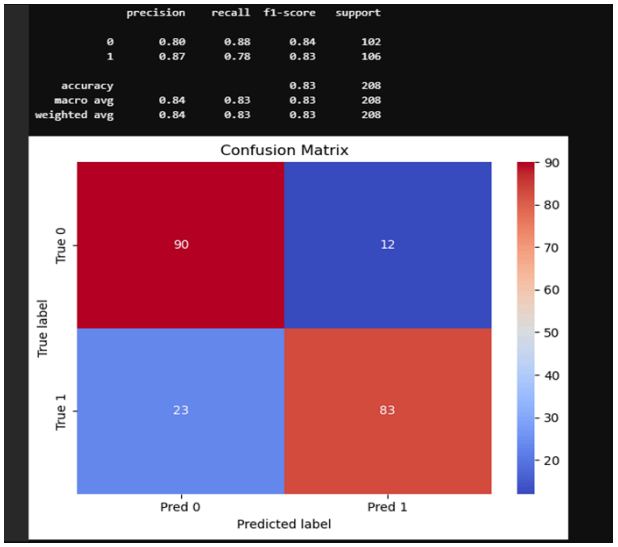
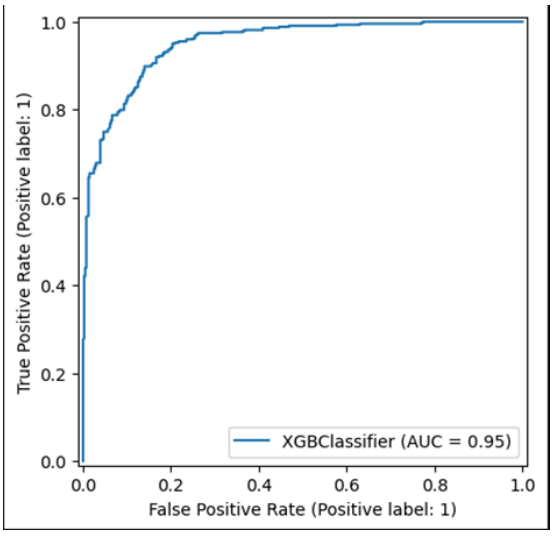
XGBoost Model Training and Validation Scores — 3km Buffer Datasets

Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
Valid_accuracy	**0.80 ± 0.05**	0.74 ± 0.03	0.74 ± 0.03	0.74 ± 0.03	0.72 ± 0.06	0.72 ± 0.02
Train_accuracy	**0.85 ± 0.01**	0.83 ± 0.01	0.82 ± 0.01	0.81 ± 0.01	0.79 ± 0.01	0.80 ± 0.01
Valid_precision	**0.83 ± 0.04**	0.75 ± 0.05	0.75 ± 0.03	0.75 ± 0.04	0.75 ± 0.07	0.72 ± 0.02
Train_precision	**0.90 ± 0.02**	0.84 ± 0.02	0.82 ± 0.01	0.82 ± 0.02	0.81 ± 0.02	0.81 ± 0.01
Valid_recall	**0.75 ± 0.07**	0.76 ± 0.05	0.75 ± 0.05	0.75 ± 0.05	0.70 ± 0.07	0.74 ± 0.07
Train_recall	**0.81 ± 0.02**	0.82 ± 0.01	0.82 ± 0.02	0.80 ± 0.01	0.76 ± 0.01	0.81 ± 0.02
Valid_f1	**0.79 ± 0.06**	0.75 ± 0.03	0.75 ± 0.03	0.75 ± 0.03	0.72 ± 0.06	0.73 ± 0.03
Train_f1	**0.85 ± 0.01**	0.83 ± 0.01	0.82 ± 0.01	0.81 ± 0.01	0.79 ± 0.01	0.81 ± 0.01

XGBoost Model Testing Scores — 3km Buffer Datasets

Metric	SS01_3K	TS01_3K	TS02_3K	TS03_3K	TS04_3K	TS05_3K
Test_accuracy	**0.83**	0.76	0.77	0.78	0.73	0.79
Test_precision	**0.84**	0.76	0.77	0.78	0.73	0.79
Test_recall	**0.83**	0.76	0.77	0.78	0.73	0.79
Test_f1	**0.83**	0.76	0.77	0.78	0.73	0.79

XGBoost performace for 3km buffered soil informed dataset

Confusion matrix	ROC-AUC
	

3. Random Forest Model Performance

Stable and Reliable: Random Forest provides consistent baseline performance across all configurations.

Key Strengths:

- Most stable performance across different datasets
- Minimal variance in results
- Good baseline model with predictable behavior
- Lower training accuracy suggests less overfitting

1km Buffer Results:

- **Validation Accuracy**: 0.68-0.76 (mean)
- **Test Performance**: Consistent across datasets (0.69-0.72)
- **Average TS Performance**: 0.71

Random Forest Model — Training and Validation Scores (1km Buffer Datasets)

Metric	SS01_1K (mean ± std)	TS01_1K (mean ± std)	TS02_1K (mean ± std)	TS03_1K (mean ± std)	TS04_1K (mean ± std)	TS05_1K (mean ± std)
Valid_accuracy	**0.76 ± 0.04**	0.70 ± 0.03	0.70 ± 0.08	0.70 ± 0.05	0.68 ± 0.02	0.70 ± 0.04
Train_accuracy	**0.81 ± 0.01**	0.74 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.73 ± 0.02	0.75 ± 0.01
Valid_precision	**0.80 ± 0.05**	0.70 ± 0.03	0.76 ± 0.12	0.73 ± 0.06	0.71 ± 0.04	0.70 ± 0.06
Train_precision	**0.86 ± 0.02**	0.75 ± 0.02	0.82 ± 0.03	0.77 ± 0.02	0.77 ± 0.02	0.75 ± 0.02

Metric	SS01_1K (mean ± std)	TS01_1K (mean ± std)	TS02_1K (mean ± std)	TS03_1K (mean ± std)	TS04_1K (mean ± std)	TS05_1K (mean ± std)
Valid_recall	**0.72 ± 0.05**	0.70 ± 0.08	0.63 ± 0.08	0.68 ± 0.09	0.64 ± 0.08	0.72 ± 0.09
Train_recall	**0.76 ± 0.02**	0.72 ± 0.01	0.66 ± 0.02	0.72 ± 0.01	0.68 ± 0.05	0.75 ± 0.02
Valid_f1	**0.75 ± 0.04**	0.70 ± 0.04	0.68 ± 0.08	0.70 ± 0.06	0.67 ± 0.04	0.71 ± 0.05
Train_f1	**0.80 ± 0.01**	0.73 ± 0.01	0.73 ± 0.02	0.74 ± 0.01	0.72 ± 0.03	0.75 ± 0.01

Random Forest Model — Testing Scores (1km Buffer Datasets)

Metric	SS01_1K	TS01_1K	TS02_1K	TS03_1K	TS04_1K	TS05_1K
Test_accuracy	**0.80**	0.72	0.70	0.72	0.72	0.70
Test_precision	**0.81**	0.72	0.70	0.72	0.72	0.70
Test_recall	**0.81**	0.72	0.70	0.72	0.72	0.69
Test_f1	**0.80**	0.72	0.69	0.72	0.72	0.69

3km Buffer Results:

- **Validation Accuracy**: 0.69-0.78 (mean)
- **Test Performance**: Slightly improved (0.70-0.75)
- **Average TS Performance**: 0.73

Random Forest Model — Training and Validation Scores (3km Buffer Datasets)

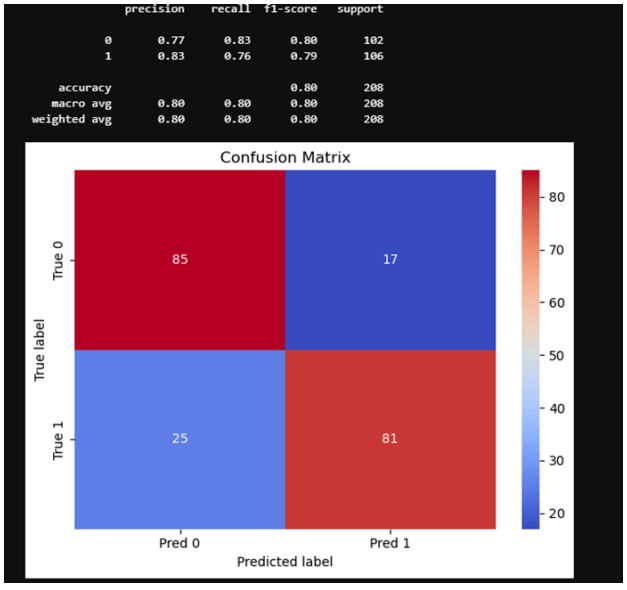
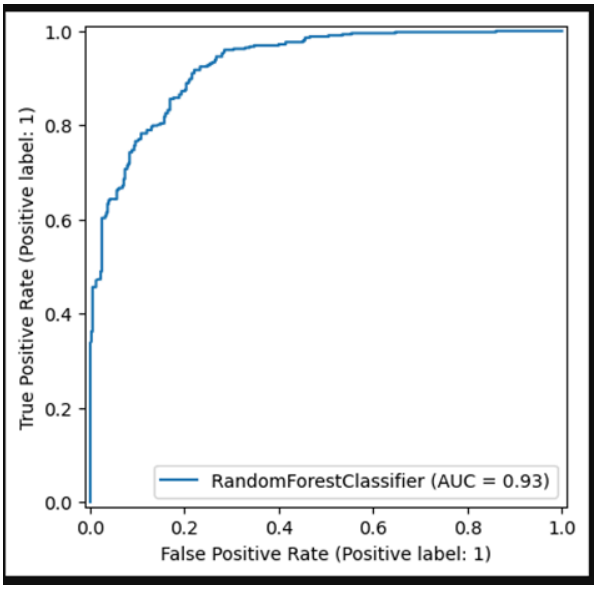
Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
Valid_accuracy	**0.78 ± 0.05**	0.69 ± 0.04	0.71 ± 0.03	0.71 ± 0.02	0.70 ± 0.06	0.69 ± 0.03
Train_accuracy	**0.83 ± 0.01**	0.74 ± 0.01	0.74 ± 0.01	0.74 ± 0.00	0.75 ± 0.00	0.75 ± 0.00
Valid_precision	**0.82 ± 0.05**	0.71 ± 0.05	0.73 ± 0.03	0.73 ± 0.03	0.75 ± 0.06	0.72 ± 0.03
Train_precision	**0.87 ± 0.01**	0.77 ± 0.02	0.77 ± 0.01	0.77 ± 0.01	0.79 ± 0.01	0.78 ± 0.01
Valid_recall	**0.73 ± 0.07**	0.67 ± 0.07	0.69 ± 0.05	0.68 ± 0.07	0.64 ± 0.08	0.66 ± 0.08
Train_recall	**0.77 ± 0.02**	0.70 ± 0.01	0.71 ± 0.02	0.71 ± 0.00	0.69 ± 0.01	0.71 ± 0.02
Valid_f1	**0.77 ± 0.06**	0.69 ± 0.05	0.71 ± 0.03	0.70 ± 0.03	0.69 ± 0.07	0.69 ± 0.05

Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
Train_f1	**0.82 ± 0.01**	0.73 ± 0.01	0.74 ± 0.01	0.74 ± 0.00	0.74 ± 0.01	0.75 ± 0.01

Random Forest Model — Testing Scores (3km Buffer Datasets)

Metric	SS01_3K	TS01_3K	TS02_3K	TS03_3K	TS04_3K	TS05_3K
Test_accuracy	**0.80**	0.71	0.74	0.75	0.70	0.74
Test_precision	**0.80**	0.71	0.74	0.75	0.70	0.75
Test_recall	**0.80**	0.71	0.74	0.75	0.70	0.74
Test_f1	**0.80**	0.71	0.74	0.75	0.70	0.74

Random Forest model performance on soil informed 3km buffered dataset

Confusion matrix	ROC-AUC
	

4. Logistic Regression Model Performance

Simplest Baseline: Logistic Regression provides the baseline linear model performance.

Key Characteristics:

- Lowest overall performance but fastest training
- Minimal overfitting (small train-validation gaps)
- Good interpretability for understanding feature importance
- Suitable for baseline comparisons

1km Buffer Results:

- **Validation Accuracy**: 0.67-0.74 (mean)

- **Test Performance**: Most variable (0.62-0.76)
- **Average TS Performance**: 0.68 (lowest among all models)

Logistic Regression Model — Training and Validation Scores (1km Buffer Datasets)

Metric	SS01_1K (mean ± std)	TS01_1K (mean ± std)	TS02_1K (mean ± std)	TS03_1K (mean ± std)	TS04_1K (mean ± std)	TS05_1K (mean ± std)
Valid_accuracy	**0.74 ± 0.05**	0.70 ± 0.04	0.67 ± 0.06	0.68 ± 0.07	0.68 ± 0.02	0.69 ± 0.06
Train_accuracy	**0.75 ± 0.02**	0.73 ± 0.01	0.72 ± 0.02	0.70 ± 0.01	0.73 ± 0.01	0.71 ± 0.02
Valid_precision	**0.79 ± 0.05**	0.72 ± 0.05	0.70 ± 0.07	0.71 ± 0.08	0.71 ± 0.06	0.70 ± 0.07
Train_precision	**0.80 ± 0.02**	0.76 ± 0.01	0.76 ± 0.03	0.73 ± 0.02	0.77 ± 0.01	0.73 ± 0.03
Valid_recall	**0.66 ± 0.09**	0.66 ± 0.07	0.62 ± 0.08	0.65 ± 0.09	0.63 ± 0.05	0.67 ± 0.08
Train_recall	**0.67 ± 0.02**	0.70 ± 0.02	0.66 ± 0.03	0.65 ± 0.03	0.67 ± 0.03	0.68 ± 0.04
Valid_f1	**0.72 ± 0.07**	0.69 ± 0.05	0.66 ± 0.07	0.67 ± 0.07	0.67 ± 0.02	0.68 ± 0.06
Train_f1	**0.73 ± 0.02**	0.73 ± 0.01	0.70 ± 0.03	0.69 ± 0.02	0.71 ± 0.02	0.70 ± 0.03

Logistic Regression Model — Testing Scores (1km Buffer Datasets)

Metric	SS01_1K	TS01_1K	TS02_1K	TS03_1K	TS04_1K	TS05_1K
Test_accuracy	**0.76**	0.71	0.68	0.66	0.71	0.62
Test_precision	**0.76**	0.71	0.69	0.66	0.71	0.62
Test_recall	**0.76**	0.71	0.68	0.66	0.71	0.62
Test_f1	**0.76**	0.71	0.68	0.66	0.71	0.62

3km Buffer Results:

- **Validation Accuracy**: 0.66-0.74 (mean)
- **Test Performance**: Improved consistency (0.69-0.75)
- **Average TS Performance**: 0.72

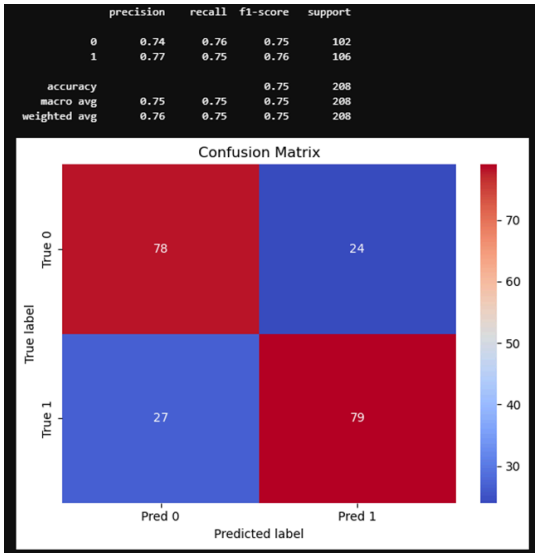
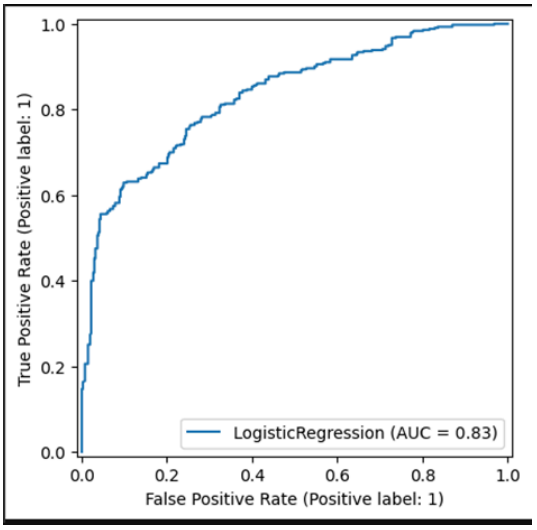
Logistic Regression Model — Training and Validation Scores (3km Buffer Datasets)

Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
Valid_accuracy	**0.74 ±**	0.70 ±	0.70 ±	0.69 ±	0.68 ±	0.66 ±

Metric	SS01_3K (mean ± std)	TS01_3K (mean ± std)	TS02_3K (mean ± std)	TS03_3K (mean ± std)	TS04_3K (mean ± std)	TS05_3K (mean ± std)
	0.06	0.04	0.03	0.01	0.05	0.04
Train_accuracy	**0.74 ± 0.02**	0.72 ± 0.02	0.71 ± 0.01	0.71 ± 0.00	0.69 ± 0.02	0.68 ± 0.01
Valid_precision	**0.78 ± 0.07**	0.72 ± 0.03	0.74 ± 0.03	0.73 ± 0.03	0.71 ± 0.04	0.69 ± 0.04
Train_precision	**0.78 ± 0.02**	0.74 ± 0.02	0.75 ± 0.01	0.75 ± 0.01	0.72 ± 0.02	0.71 ± 0.01
Valid_recall	**0.68 ± 0.07**	0.68 ± 0.06	0.66 ± 0.04	0.65 ± 0.05	0.66 ± 0.08	0.63 ± 0.08
Train_recall	**0.68 ± 0.03**	0.69 ± 0.03	0.66 ± 0.02	0.66 ± 0.01	0.66 ± 0.02	0.64 ± 0.01
Valid_f1	**0.73 ± 0.07**	0.70 ± 0.04	0.70 ± 0.03	0.68 ± 0.02	0.68 ± 0.06	0.65 ± 0.05
Train_f1	**0.73 ± 0.02**	0.72 ± 0.02	0.70 ± 0.01	0.70 ± 0.00	0.69 ± 0.02	0.68 ± 0.01

Logistic Regression Model — Testing Scores (3km Buffer Datasets)

Metric	SS01_3K	TS01_3K	TS02_3K	TS03_3K	TS04_3K	TS05_3K
Test_accuracy	**0.75**	0.69	0.74	0.75	0.70	0.74
Test_precision	**0.75**	0.69	0.74	0.75	0.70	0.74
Test_recall	**0.75**	0.69	0.74	0.75	0.70	0.74
Test_f1	**0.75**	0.69	0.74	0.75	0.70	0.74

Confusion matrix	ROC-AUC
	

Comparative Analysis

Model performace for 1km buffer datasets

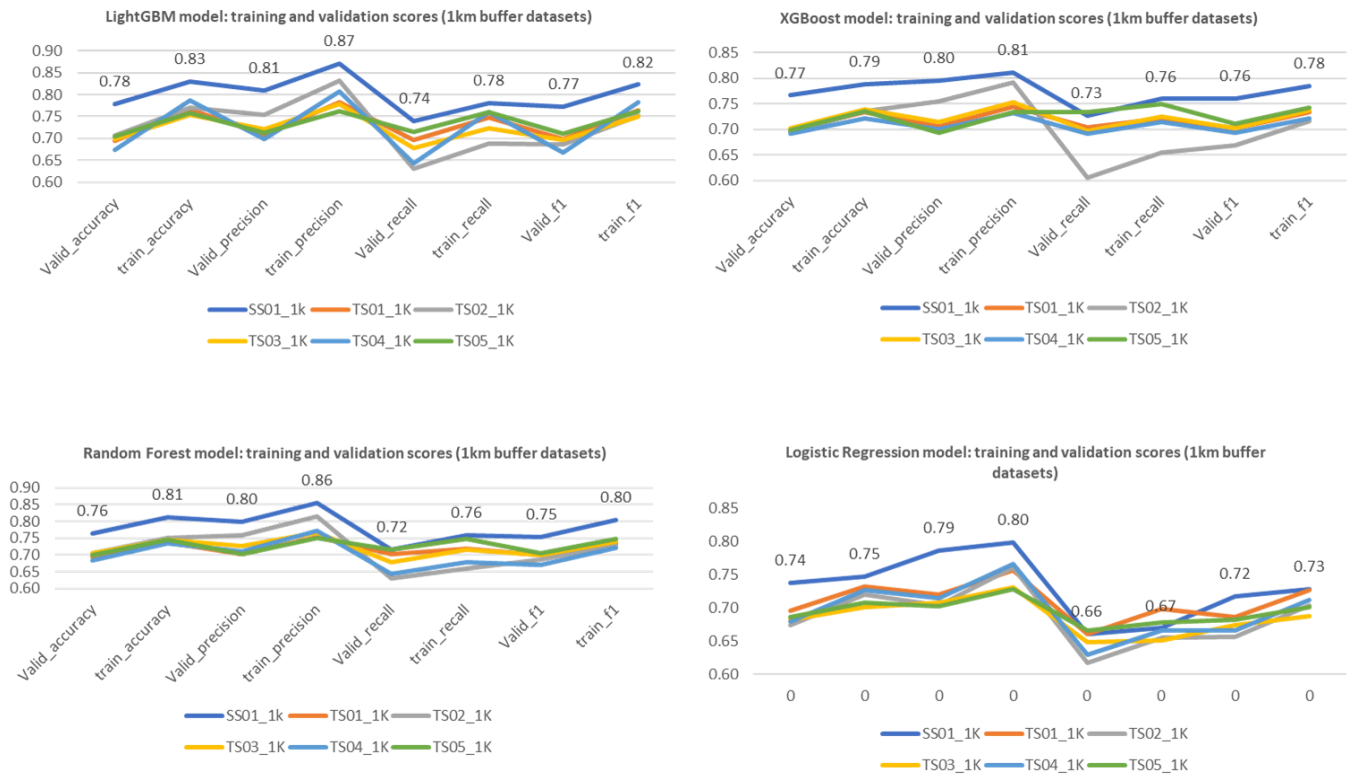


Fig. 1: Comparitive analysis of model performace 1km buffer.png

Comparative analysis of model performace 1km buffer.png

Comparitive analysis of model performace 1km buffer

Model performance for 3 km buffer datasets

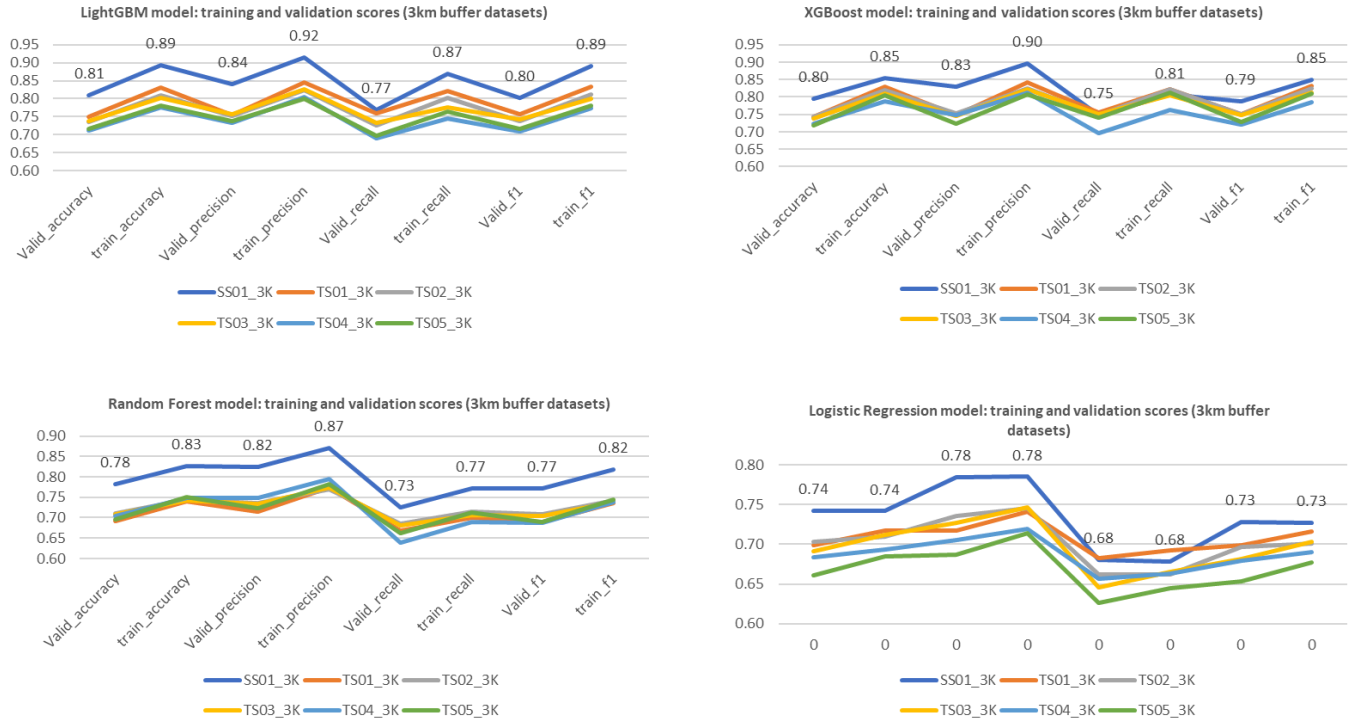


Fig. 2: Comparitive analysis of model performace 3km buffer.png

Comparative analysis of model performance 3km buffer

Overall Model Ranking:

1. **LightGBM** - Best overall performance, highest accuracy and F1-scores

2. **XGBoost** - Strong second choice with good generalization
3. **Random Forest** - Reliable baseline with consistent results
4. **Logistic Regression** - Simple baseline with adequate performance

---

Buffer Zone Impact:

- **3km buffers generally outperform 1km buffers** across all models
- Larger buffers provide more training data and better spatial context
- LightGBM shows the most significant improvement (0.71 to 0.75 average TS performance)

Dataset-Specific Insights:

- **SS01 datasets** consistently achieve highest performance, indicating the value of domain knowledge in negative sampling
- **TS datasets** show more variability but maintain reasonable consistency
- Random negative sampling (TS) provides robust cross-validation of model performance

Recommendations

Primary Recommendation:

Deploy LightGBM with 3km buffer using SS01 negative sampling strategy for production mineral prospectivity mapping.

Supporting Rationale:

1. **Highest Accuracy**: 0.85 test accuracy on SS01_3K dataset
2. **Balanced Performance**: Excellent precision-recall balance
3. **Consistent Results**: Strong performance across different validation strategies
4. **Scalability**: Efficient training and prediction capabilities

Secondary Options:

- **XGBoost with 3km buffer** as a secondary validation model
- **Random Forest** for ensemble approaches and feature importance analysis
- **Logistic Regression** for interpretable baseline comparisons

Implementation Strategy:

1. Prioritize soil-informed negative sampling (SS01 approach)
2. Use 3km buffer zones for optimal spatial context
3. Implement ensemble methods combining top-performing models
4. Validate results using multiple TS datasets for robustness testing

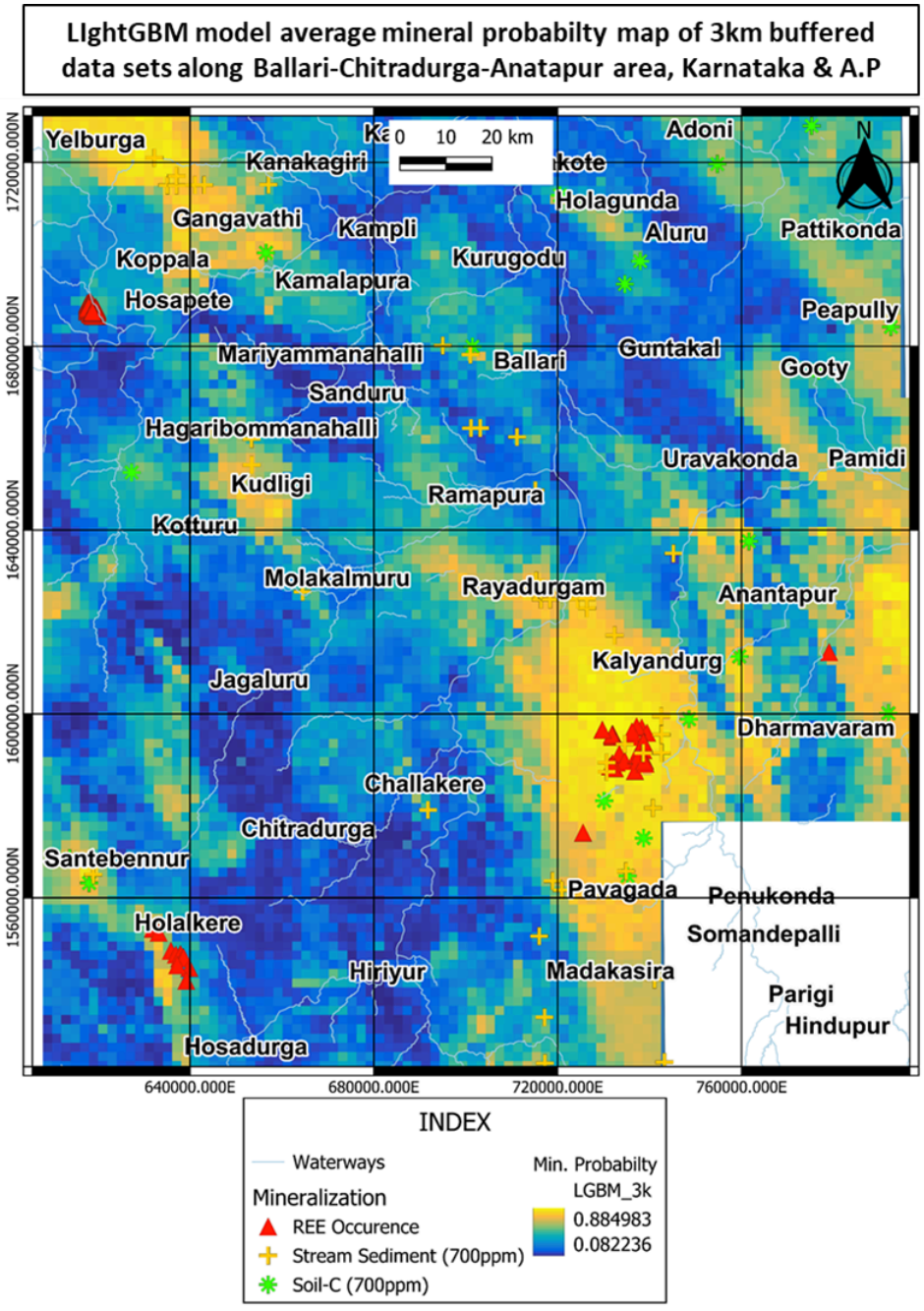
The analysis demonstrates that gradient boosting methods (LightGBM and XGBoost) significantly outperform traditional methods for REE mineral prospectivity mapping. The superior performance of soil-informed negative sampling highlights the importance of domain expertise in geological machine learning applications. The consistent

improvement with larger buffer zones suggests that spatial context is crucial for accurate mineral prospectivity prediction.

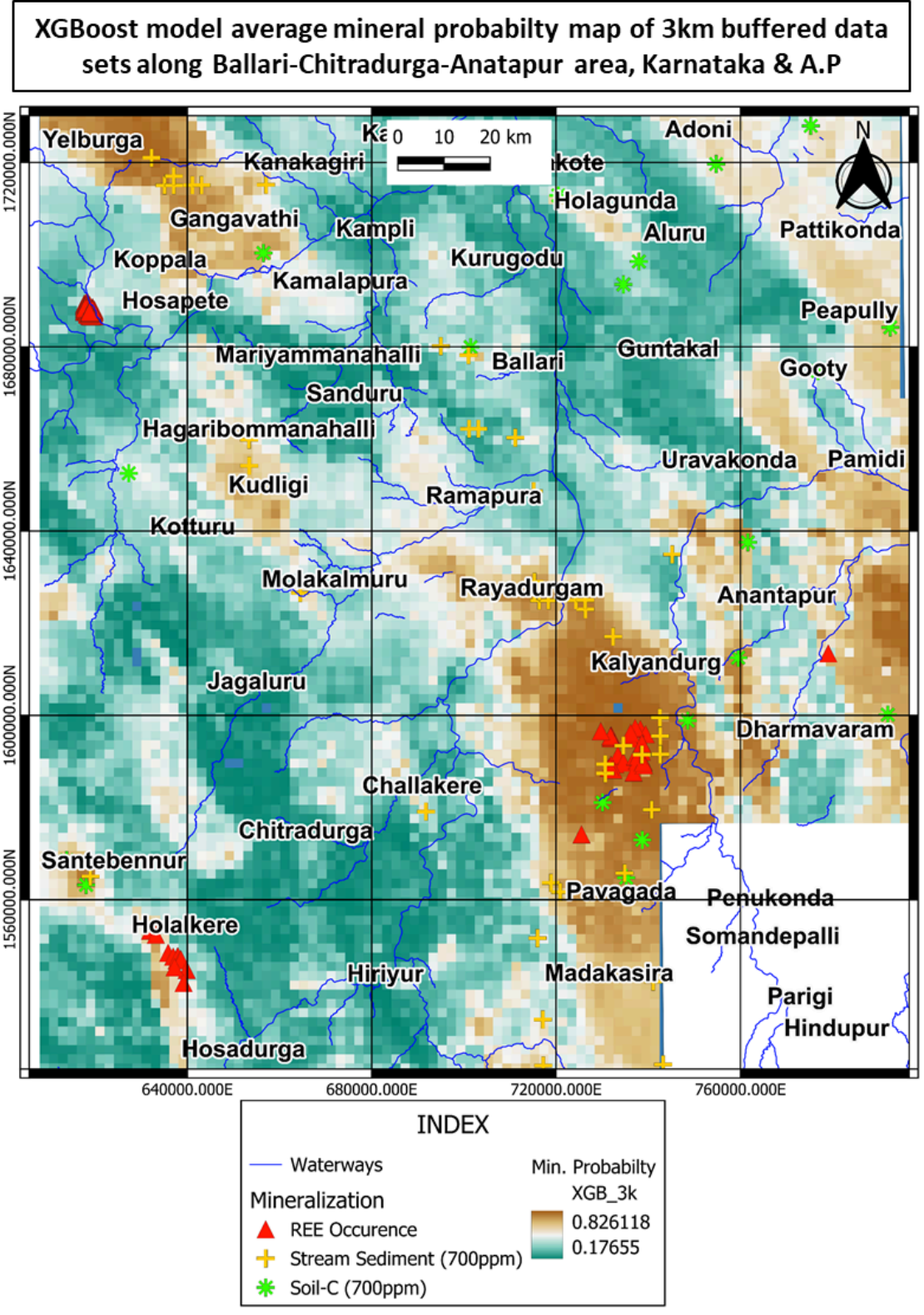
LightGBM emerges as the clear winner for this application, offering the best combination of accuracy, consistency, and practical performance for REE mineral exploration efforts.

IX. Mineral prospectivity maps

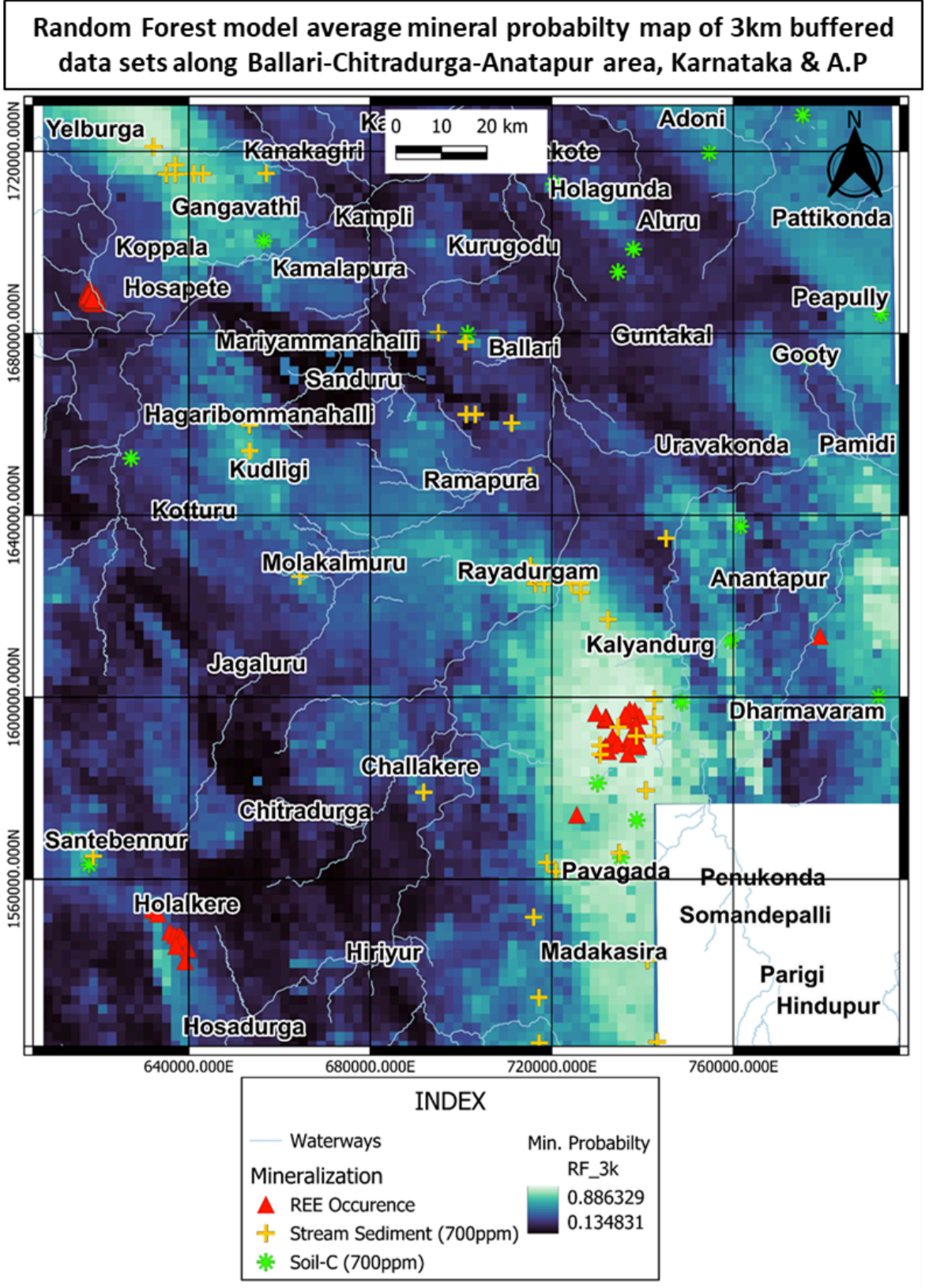
To visualize and compare the predictive capabilities of individual models, mineral prospectivity maps were generated using the 3 km buffer training datasets for each of the four algorithms—LightGBM, XGBoost, Random Forest, and Logistic Regression. These maps represent the spatial distribution of prospectivity scores derived from the averaged probability outputs of multiple cross-validated models. By highlighting areas of higher predicted likelihood for REE mineralization, each map offers a unique perspective on potential targets, reflecting the strengths and spatial sensitivity of the underlying model. These visualizations also serve as the foundation for ensembling strategies and further spatial analysis.



Mineral prospectivity map using LightGBM model



Mineral Prospectivity maps using XGBoost models



Mineral prospectivity map using Random Forest model

X. Ensembling Strategy and Performance Evaluation

To enhance the robustness and generalization of mineral prospectivity modeling using the SS01_3K dataset (3 km buffer), two ensemble strategies were implemented:

1. **Simple Averaging Ensemble** – An unweighted average of the predicted probabilities from the top three base models (LightGBM, XGBoost, and Random Forest).
2. **Stacking Ensemble with Gradient Boosting Meta-Learner** – A more sophisticated approach using a gradient boosting classifier as the meta-learner trained on calibrated probabilistic outputs from base models.

The rationale behind these strategies was to capture the complementary strengths of different machine learning models and mitigate individual weaknesses.

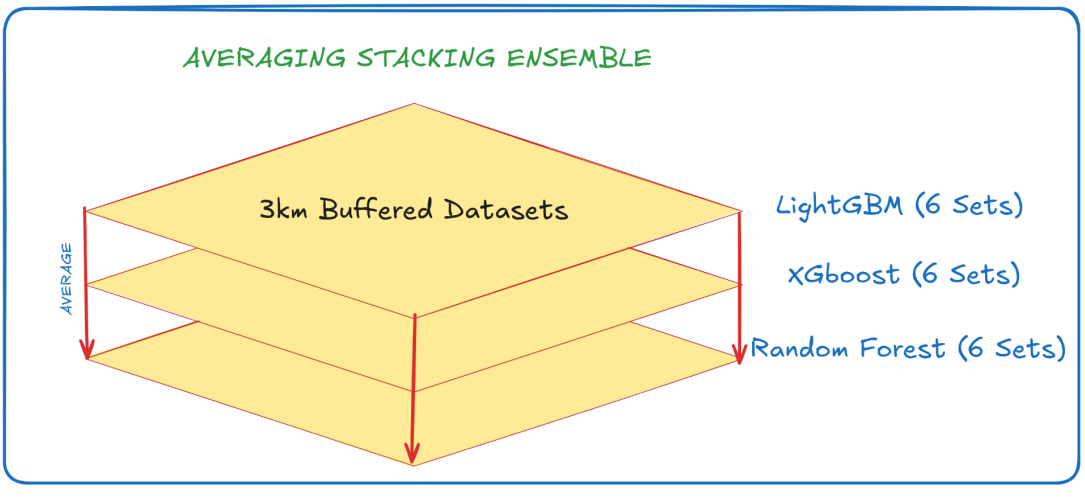
Base Learners and Their Performance

Four models were used as base learners:

Model	Test Accuracy	F1 Score	ROC-AUC	Notes
LightGBM	0.85	0.85	**0.98**	Best individual model
XGBoost	0.83	0.83	0.95	Strong second-best
Random Forest	0.80	0.80	0.93	Balanced, moderately complex
Logistic Regr.	0.75	0.75	0.83	Linear model, interpretable

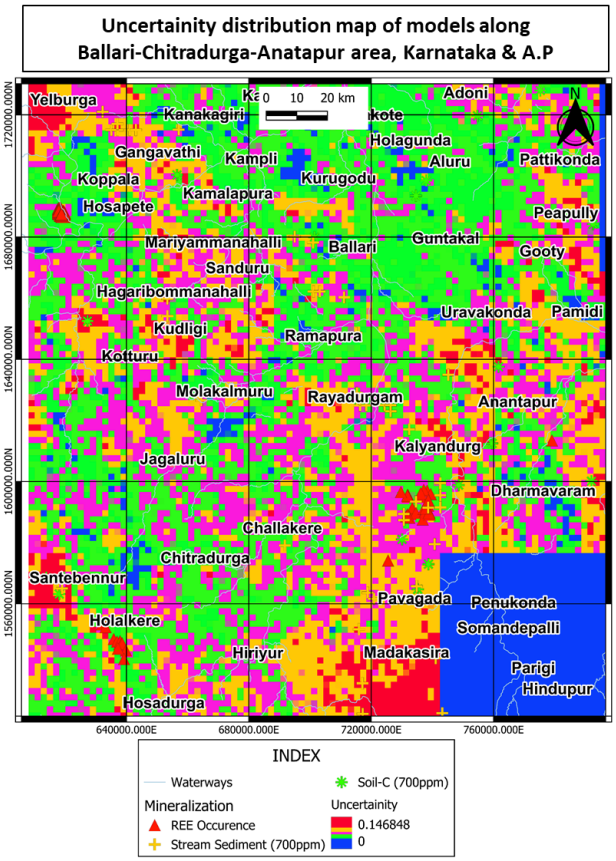
LightGBM achieved the highest ROC-AUC and F1-score, indicating strong performance in both precision and recall.

Ensemble Strategy I: Simple Averaging of Probabilities



In this approach, **no binary outputs were considered**; instead, the **average of the probabilistic predictions** from LightGBM, XGBoost, and Random Forest was computed for each instance. Final classification was done by thresholding this averaged probability.

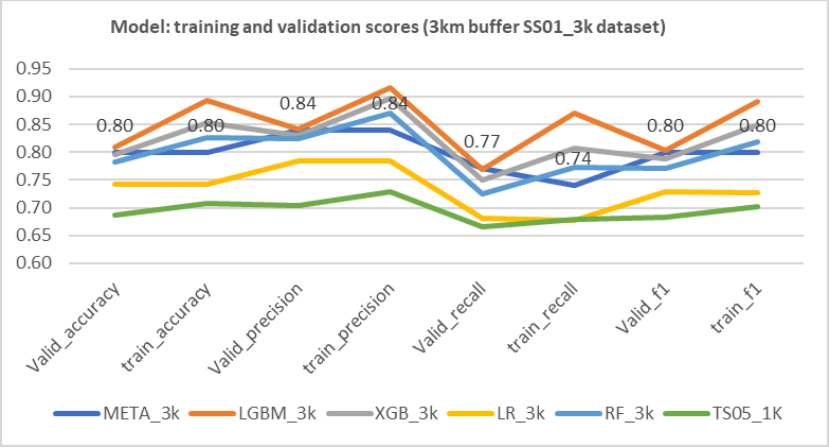
Although no specific test metrics were logged for this ensemble, preliminary inspection indicated it achieved **slightly smoother predictions** and **reduced variance**. In the simple averaging ensemble strategy, uncertainty quantification was performed using the **standard deviation of the predicted probabilities across 48 individual models**, which comprised multiple cross-validation runs of the four base learners—LightGBM, XGBoost, Logistic Regression and Random Forest. By capturing the **spread in predicted probabilities** at each pixel or sample location, this method provided a **measure of model disagreement**, with higher standard deviation indicating greater uncertainty. This approach allowed for **spatially-resolved uncertainty estimation**, helping identify areas where the ensemble was less confident in its prospectivity predictions. Such probabilistic dispersion is particularly valuable in mineral exploration, where decision-making under uncertainty is common, enabling geoscientists to prioritize follow-up actions not only based on prospectivity but also based on **confidence levels in model predictions**.



Ensemble Strategy II: Stacking with Gradient Boosting Meta-Learner

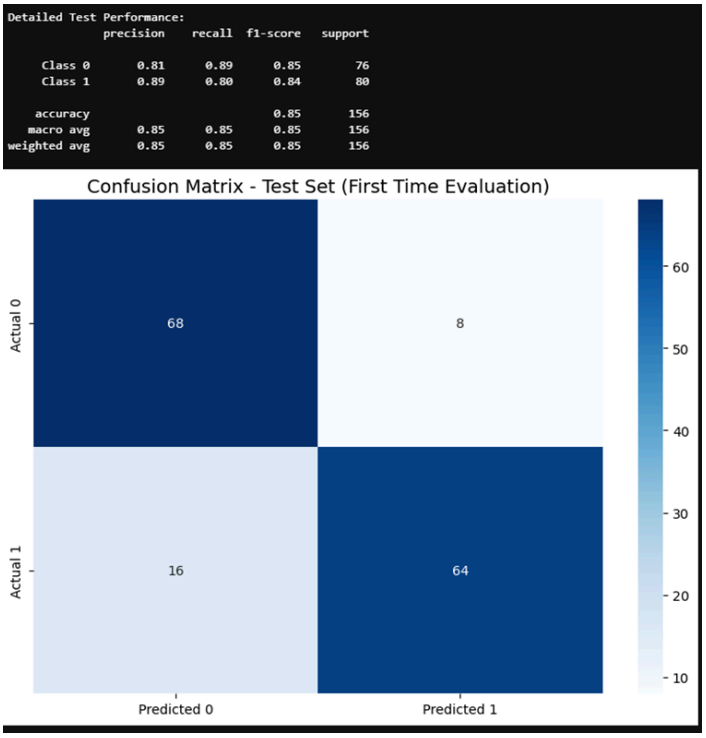
A more powerful stacking ensemble was implemented using a **Gradient Boosting classifier as the meta-learner**. The key features of this strategy include:

- **Inputs to meta-learner**: Calibrated probabilistic outputs from LGBM, XGB, RF, and LR.
- **Training procedure**: Nested cross-validation to avoid data leakage.
- **Target**: Probablistic predictions of REE prospectivity.
- **Calibration**: Ensures the meta-learner can effectively interpret probability outputs from heterogeneous models.



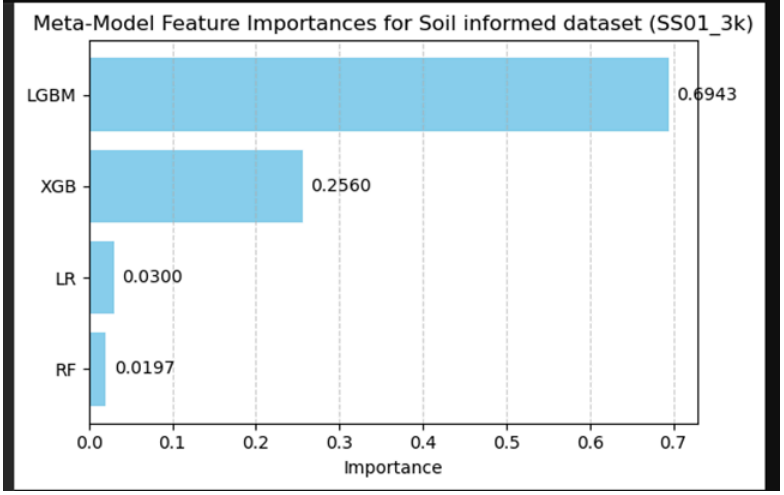
Comparision of metamodel with base learners

Metric	Stacking (Meta)	LGBM	XGBoost	RF	LR
Test Accuracy	**0.85**	0.85	0.83	0.80	0.75
Test F1	**0.85**	0.85	0.83	0.80	0.75
ROC-AUC	**0.872**	**0.98**	0.95	0.93	0.83



➤ Meta-Learner Feature Importance:

The feature importance within the meta-learner (Gradient Boosting) reveals:



Feature importance of Meta learner

Feature importance of Meta learner

Feature importance of Meta learner

- **LightGBM contributed the most**, confirming its dominance as the strongest signal.
- **XGBoost made moderate contributions**, complementing LGBM with different splits.
- **Random Forest and Logistic Regression added minor but non-zero importance**, showing some unique information.

Interpretation and Discussion

The stacking model demonstrated **robust and consistent performance**, matching the accuracy and F1-score of the best base model (LGBM) while offering the added benefits of:

- **Reduced overfitting** (train accuracy = 0.80 vs. LGBM's 0.89)
- **Enhanced stability across folds** due to nested cross-validation
- **Aggregated predictive power** from diverse learners

However, the ROC-AUC of the stacked model (0.872) remained **lower than the best-performing LGBM** (0.98), suggesting that the meta-learner, while stable, does **not surpass the individual discriminatory strength of LGBM**.

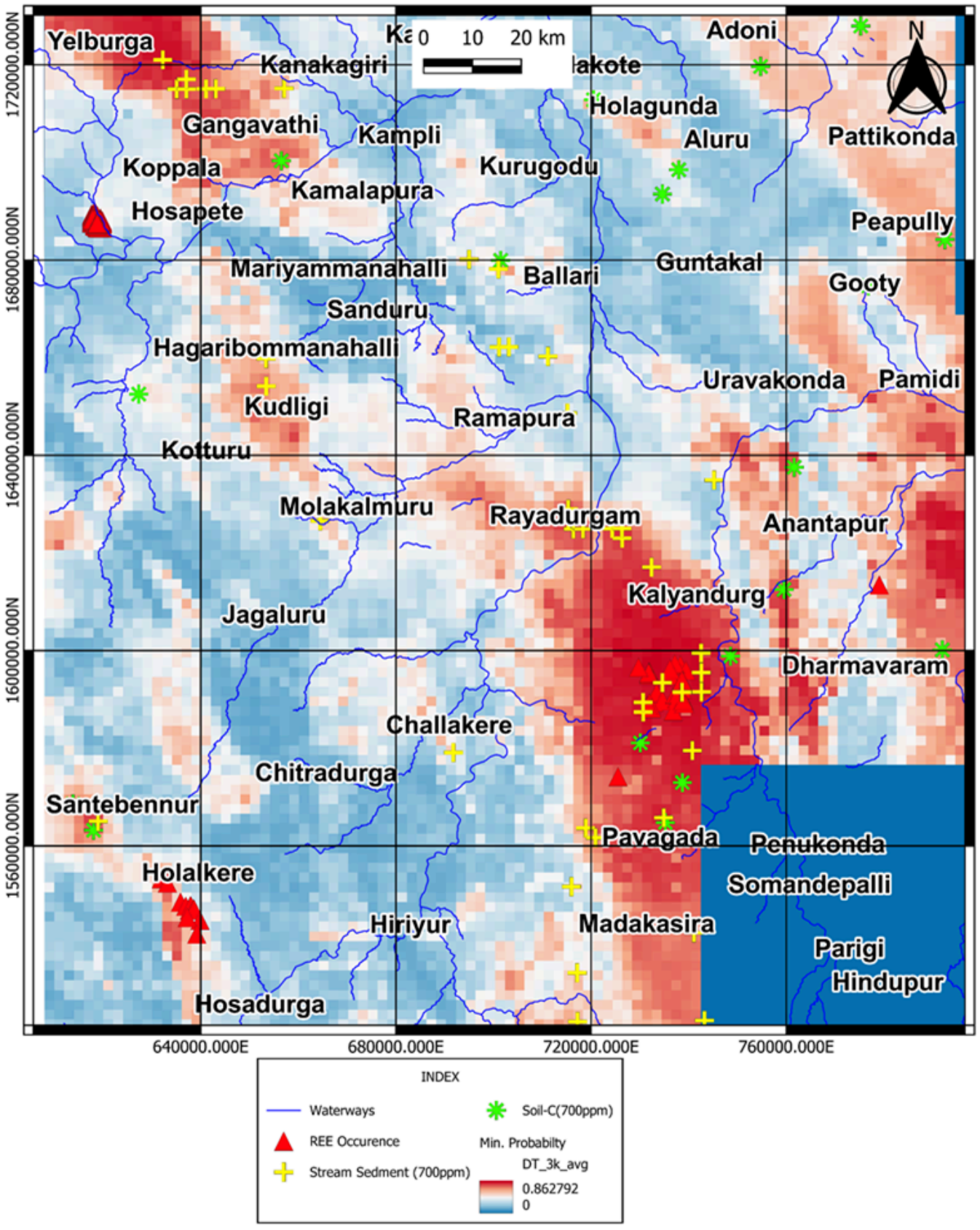
The ensembling strategies successfully integrated multiple predictive signals. The key takeaways:

- **Plain averaging** is simple and slightly smooths predictions but offers **little gain** over individual models.
- **Stacking with a Gradient Boosting meta-learner** provides a **more structured and generalizable model**, with strong balance across all metrics.
- While **LightGBM remains the top individual model**, stacking achieves comparable performance with **reduced variance and better generalization**—making it a viable candidate for operational deployment.

Future improvements could explore **weighted ensembling**, **meta-feature integration**, or **spatial cross-validation**, especially for high-resolution prospectivity modeling.

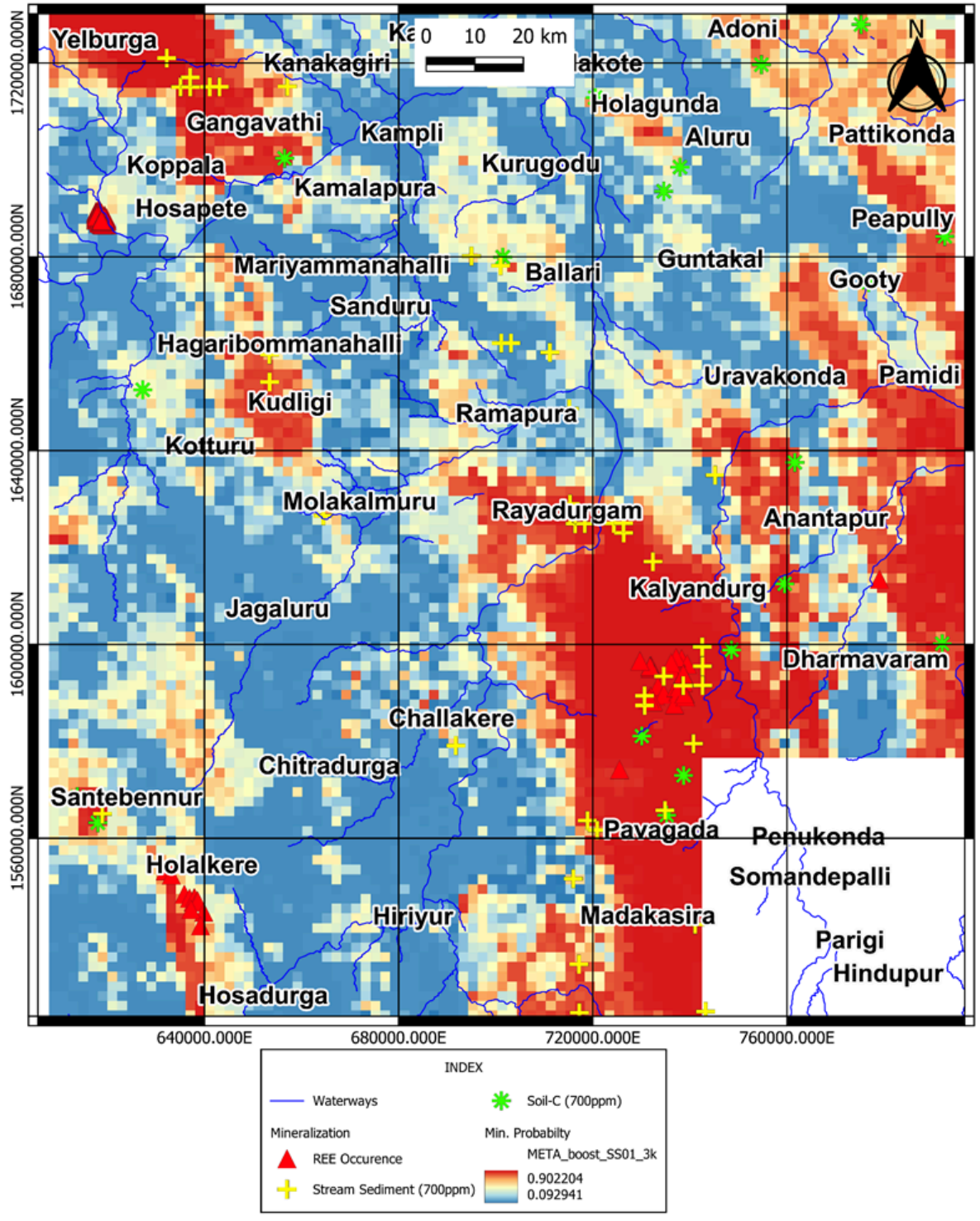
XI. Average Ensemble REE mineral probabilty maps

3 Model average (LGBM, XGB,RF) mineral probabilty map of 3km buffered data sets along Ballari-Chitradurga-Anatapur area, Karnataka & A.P



XII. Stacked ensemble MetaLearner (Gradient Boost) probabilty map

Stacked ensemble meta learner mineral probabilty map of 3km buffered data sets along Ballari-Chitradurga-Anatapur area, Karnataka & A.P



XIII. Synthesis of REE Prospectivity Targets Using Ensemble Maps

The synthesis of Rare Earth Element (REE) prospectivity targets represents a critical phase in mineral exploration, where multiple machine learning models are integrated to identify areas with the highest potential for REE mineralization. This chapter outlines the systematic approach employed to synthesize prospectivity targets through ensemble mapping techniques, incorporating both predictive probability and uncertainty quantification to establish a hierarchical target classification system.

Ensemble Map Generation

The foundation of the target synthesis process relied on ensemble maps created through two distinct approaches:

Simple Averaging Approach: Four individual machine learning models were employed: Light Gradient Boosting Machine (LGBM), Random Forest (RF), Extreme Gradient Boosting (XGB), and a metalearner algorithm. The simple averaging ensemble was constructed by computing the arithmetic mean of probability outputs from these four models across the study area. This approach provides equal weighting to each model's contribution, assuming comparable model performance and reliability.

Metalearner Approach: A sophisticated metalearner algorithm was implemented as an alternative ensemble strategy. The Metalearner was trained to optimize the combination of base model predictions, potentially capturing non-linear relationships between model outputs and identifying optimal weighting schemes that may not be apparent through simple averaging.

Uncertainty Quantification

Uncertainty assessment was conducted using a comprehensive statistical approach involving 48 individual maps generated through bootstrap sampling and cross-validation procedures. The standard deviation of predictions across these 48 maps served as the uncertainty metric, providing a robust measure of model confidence at each spatial location. This approach enables the quantification of epistemic uncertainty arising from model variability and data limitations.

Target Classification Framework

The synthesis process employed a dual-threshold system incorporating both probability and uncertainty criteria to establish three distinct target categories:

Probability Thresholding

- **Threshold for Stacking ensemble map**: 0.87 probability for high-confidence targets
- **Threshold for Averaging ensemble map**: 0.80 probability for simple averaging map comparison

Uncertainty Categories

Three uncertainty thresholds were established based on standard deviation values for stacking ensemble:

- **High confidence**: $\sigma < 0.07$
- **Moderate confidence**: $\sigma < 0.10$

- **Lower confidence**: σ < 0.12

Target Classification Results

Priority Targets (Prospective with High Confidence)

Priority targets were defined as areas exhibiting **probability values ≥ 0.87** combined with **uncertainty values < 0.07**. These targets represent locations where multiple models demonstrate strong consensus regarding REE prospectivity potential. The stringent uncertainty threshold ensures that predictions in these areas are supported by consistent model agreement, minimizing the risk of false positives. Priority targets warrant immediate detailed exploration activities including geochemical sampling, geophysical surveys, and potential drilling programs.

Potential Targets (Prospective with Moderate Confidence)

Potential targets encompass areas with **probability values ≥ 0.87** but u**ncertainty values between 0.07 and 0.10.** While these locations demonstrate significant prospectivity potential, the moderate uncertainty levels indicate some variability in model predictions. These targets represent secondary exploration opportunities that should be systematically evaluated following initial assessment of priority targets. Additional data collection and model refinement may help reduce uncertainty levels in these areas.

Exploratory Targets (Prospective with Lower Confidence)

Exploratory targets include areas meeting the **probability threshold of ≥ 0.87** but exhibiting uncertainty values **between 0.10 and 0.12**. These targets represent regions of interest that require cautious evaluation due to higher model uncertainty. They may serve as targets for reconnaissance-level exploration activities or for future consideration as additional geological and geochemical data become available to reduce uncertainty levels.

Comparative Analysis of Ensemble Approaches

Simple Averaging vs. Metalearner Performance

The comparison between simple averaging and metalearner ensemble approaches revealed distinct characteristics in target area delineation. The metalearner approach consistently identified larger prospective areas compared to simple averaging, suggesting its ability to capture subtle patterns and relationships that may be averaged out in the simple approach.

Simple Averaging Characteristics:

- Conservative target delineation
- Balanced contribution from all base models
- Probability threshold of **0.80** identified broader prospective regions
- More stable uncertainty estimates due to equal weighting

Metalearner Characteristics:

- Expanded prospective area identification
- Optimized model combination based on performance metrics
- Enhanced sensitivity to complex geological relationships
- Potential for identifying previously overlooked prospective zones

Spatial Distribution Analysis

The Metalearner approach demonstrated **superior performance in identifying elongated prospective zones** aligned with known geological structures, suggesting improved geological realism in target delineation. The expanded area coverage provided by the metalearner may capture transitional zones and structural extensions that simple averaging might underestimate.

Interpretation and Geological Significance

Priority Target Interpretation

Priority targets predominantly clustered along various granites within the study area, indicating that the model has effectively generalized a pegmatite-based targeting approach for REE mineralization within granitic terranes. This clustering pattern reflects the model's recognition of granite-hosted pegmatite systems as the primary REE mineralization style in the region.

Notably, syenite-based mineralization appears to have been underrepresented in the model predictions, likely due to the limited extent of the Koppala deposit and restricted syenite exposures within the training dataset. This limitation highlights the challenge of capturing less common mineralization styles when training data is dominated by more prevalent geological settings.

An intriguing observation is that stream sediment and soil C-horizon data collected in proximity to the established Koppala deposit do not exhibit highly anomalous TREE values, despite the known mineralization. Several explanations may account for this apparent discrepancy:

Sampling Grid Limitations: The coarse (grid spacing) regional geochemical sampling program may have failed to adequately capture the localized geochemical signature of the Koppala deposit. REE mineralization often exhibits sharp geochemical gradients, requiring dense sampling networks to effectively delineate anomalous zones.

Geomorphological Factors: The specific geomorphological setting of the Koppal area may influence dispersion patterns, with mineralization signatures being either diluted by extensive sediment transport or concentrated in areas not captured by the current sampling design.

Weathering and Mobility Considerations: The differential mobility of REE under tropical weathering conditions may result in complex secondary dispersion patterns that are not effectively captured by standard soil and stream sediment sampling protocols.

The high-confidence classification of identified targets reflects strong model consensus regarding the presence of key geological indicators including:

- Favorable granite-hosted lithologies
- Structural controls promoting pegmatite emplacement
- Geochemical signatures consistent with REE enrichment
- Proximity to known pegmatite occurrences

Identified Target Areas: Geological Context and Validation

A total of **6** targets were marked for ground validation and checking:

Target 1: Rayadurgam-Kalyanadurg Target

The Rayadurgam-Kalyanadurg target represents a high-priority prospective zone located at the critical contact between the Closepet Granite and Peninsular Gneissic Complex (PGC-II). The target area is characterized by alkali feldspar granite hosting dense northwest-southeast trending dolerite intrusions and associated quartz veining systems. This geological setting creates favorable structural conditions for REE concentration through hydrothermal processes.

The target's significance is enhanced by its position as an extensional zone to the known allanite-rich pegmatite mineralization at Obagnapalli, suggesting a potential genetic relationship and shared mineralizing system. Geochemical validation is provided by both stream sediment and soil geochemistry data. Stream sediment samples (n=6) from the Rayadurgam extension demonstrate consistently elevated Total Rare Earth Element (TREE) concentrations ranging from 857-1179 ppm, substantially above regional background levels and confirming the presence of REE enrichment in the catchment areas.

Target 2: Gangavathi-Yelburga Target

The Gangavathi-Yelburga target occurs at another strategically significant contact zone between the Closepet Granite and PGC-II, characterized by pink porphyritic granite lithologies. The area exhibits dense dolerite intrusion patterns with both northwest-southeast and northeast-southwest orientations, creating complex structural networks conducive to fluid flow and mineralization.

This target's proximity to the established Koppal REE deposit provides strong geological analogy and validates the prospectivity model's effectiveness in identifying similar geological environments. Geochemical support comes from soil C-horizon samples (n=2) showing TREE concentrations of 769-941 ppm, while stream sediment data (n=7) reveals a broader range from 734-4198 ppm, with the higher values indicating significant REE enrichment processes. The wide range in stream sediment values suggests heterogeneous distribution typical of primary REE mineralization.

Target 3: Gooty-Peapally Target

The Gooty-Peapally target is distinguished by patches of pink hornblende-biotite granite within the grey biotite gneiss of PGC-II. Its proximity to the Jonnagiri schist belt introduces additional complexity through potential contact metamorphism and associated hydrothermal activity. Dolerite vein intrusions provide structural pathways for mineralizing fluids.

Soil C-horizon geochemistry (n=2) indicates TREE concentrations of 756-1546 ppm, demonstrating significant REE enrichment in the weathering profile. The variability in these values suggests localized enrichment zones that warrant detailed follow-up investigation.

Target 4: Dharamavaram-Uravakonda Target

This target features a distinctive linear grey biotite granite of PGC-II in contact with the Ramagiri-Penakacherla greenstone belt. The granite-greenstone contact represents a fundamental lithological boundary often associated with hydrothermal activity and mineralization. Dolerite emplacement along the contact further enhances structural permeability.

Limited soil C-horizon data (n=1) shows TREE content of 890 ppm, providing preliminary geochemical validation. The linear nature of this target along the greenstone contact suggests potential for extensive mineralization

requiring systematic exploration along strike.

Target 5 : Anantapur-Pamidi Target

The Anantapur-Pamidi target exhibits patches of pink hornblende-biotite granite within grey biotite gneiss of PGC-II, similar to the Gooty-Peapally setting. Multiple dolerite dyke orientations (northwest-southeast and northeast-southwest) create favorable structural complexity for REE concentration.

Soil C-horizon geochemistry (n=1) from south of the area indicates TREE content of 797 ppm, confirming geochemical anomalism. The patchy nature of the granite occurrences suggests localized hydrothermal centers requiring detailed geological mapping.

Target 6: Santebennur-Hosadurga Target

The Santebennur-Hosadurga target is characterized by linear grey granite within migmatitic gneiss of PGC-I, representing a different crustal domain from other targets. Its proximity to the Chitradurga schist belt and known Ghatoshalli REE and Rare Metal-Rare Earth (RMRE) occurrences provides strong geological precedent for REE mineralization potential.

Geochemical validation comes from soil C-horizon samples at Santebennur (n=2) showing consistent TREE values of 702-705 ppm, while stream sediment data (n=1) indicates 907 ppm TREE. The consistency in soil values suggests uniform enrichment processes, while the elevated stream sediment value confirms catchment-scale REE dispersion.

Uncertainty Distribution Patterns

Spatial analysis of uncertainty patterns revealed important insights into model reliability across different geological domains. Lower uncertainty values were consistently observed in areas with:

- Dense geochemical sampling coverage
- Well-characterized geological units
- Consistent geophysical signatures
- Proximity to training data locations

Higher uncertainty regions typically corresponded to:

- Areas with sparse data coverage
- Complex geological transitions
- Limited historical exploration activity
- Edge effects near study area boundaries

The ensemble mapping approach provides a robust framework for REE prospectivity assessment, with the hierarchical target classification enabling efficient resource allocation for exploration activities. The integration of uncertainty quantification adds a critical dimension to target evaluation, allowing for risk-informed decision making.

Future Enhancements:

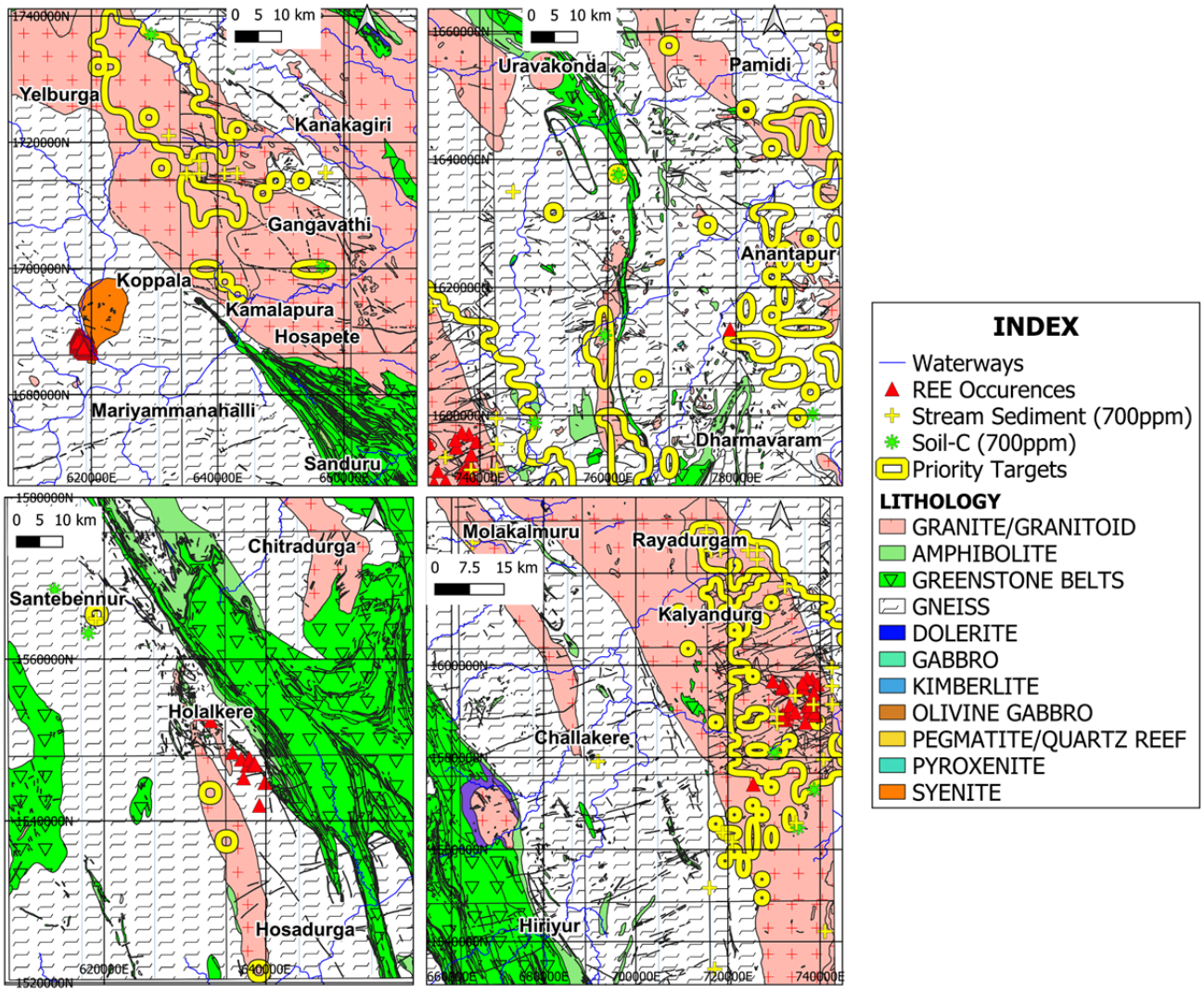
1. Incorporate additional base models to improve ensemble diversity
2. Implement dynamic uncertainty thresholds based on local geological complexity

3. Integrate economic and logistical factors into target prioritization
4. Develop temporal updating protocols as new data becomes available

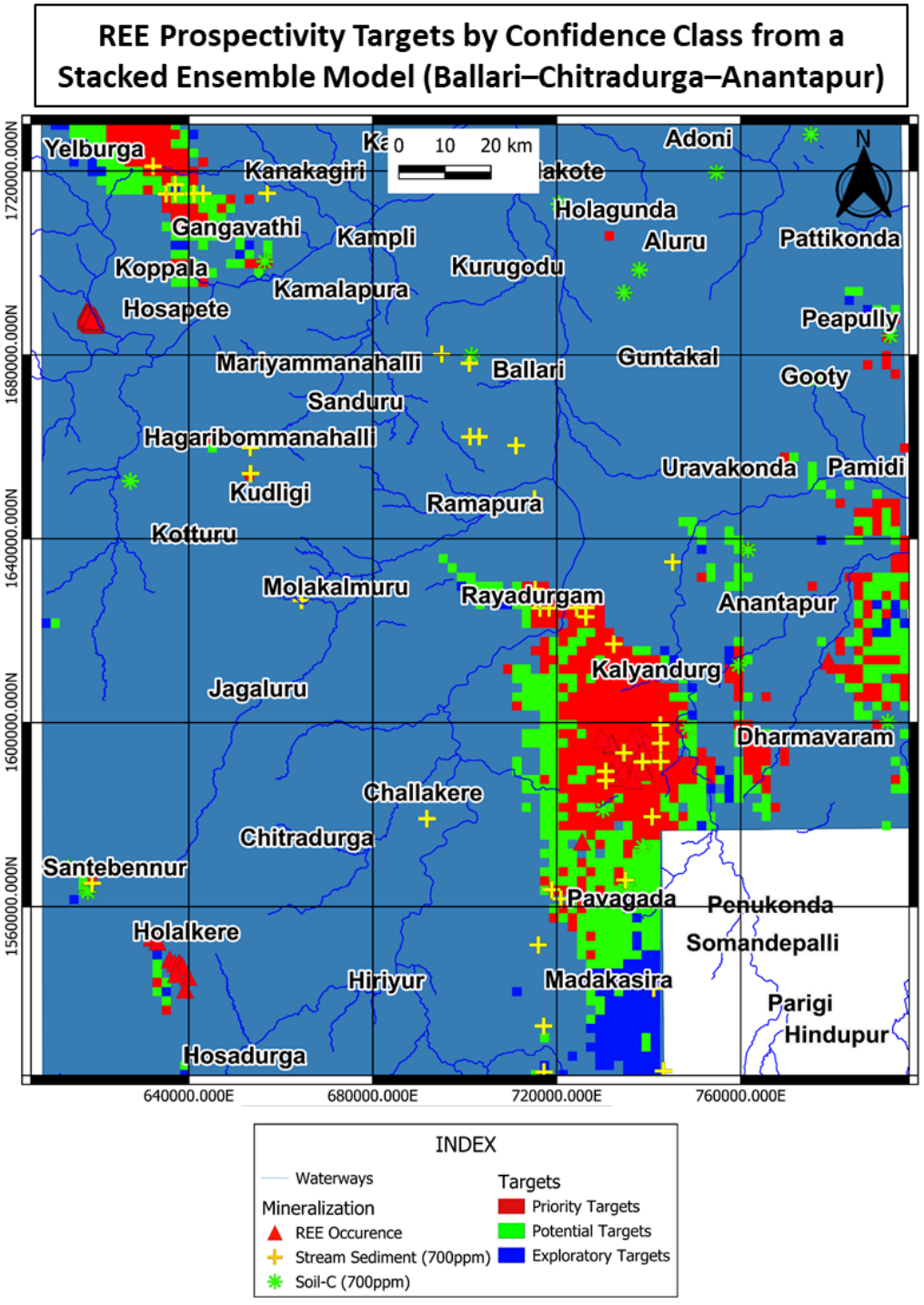
The implementation of this ensemble-based target synthesis approach demonstrates the value of combining multiple machine learning algorithms with uncertainty quantification to enhance REE exploration effectiveness while managing exploration risk through systematic target classification.

Bhukosh map showing REE priority targets

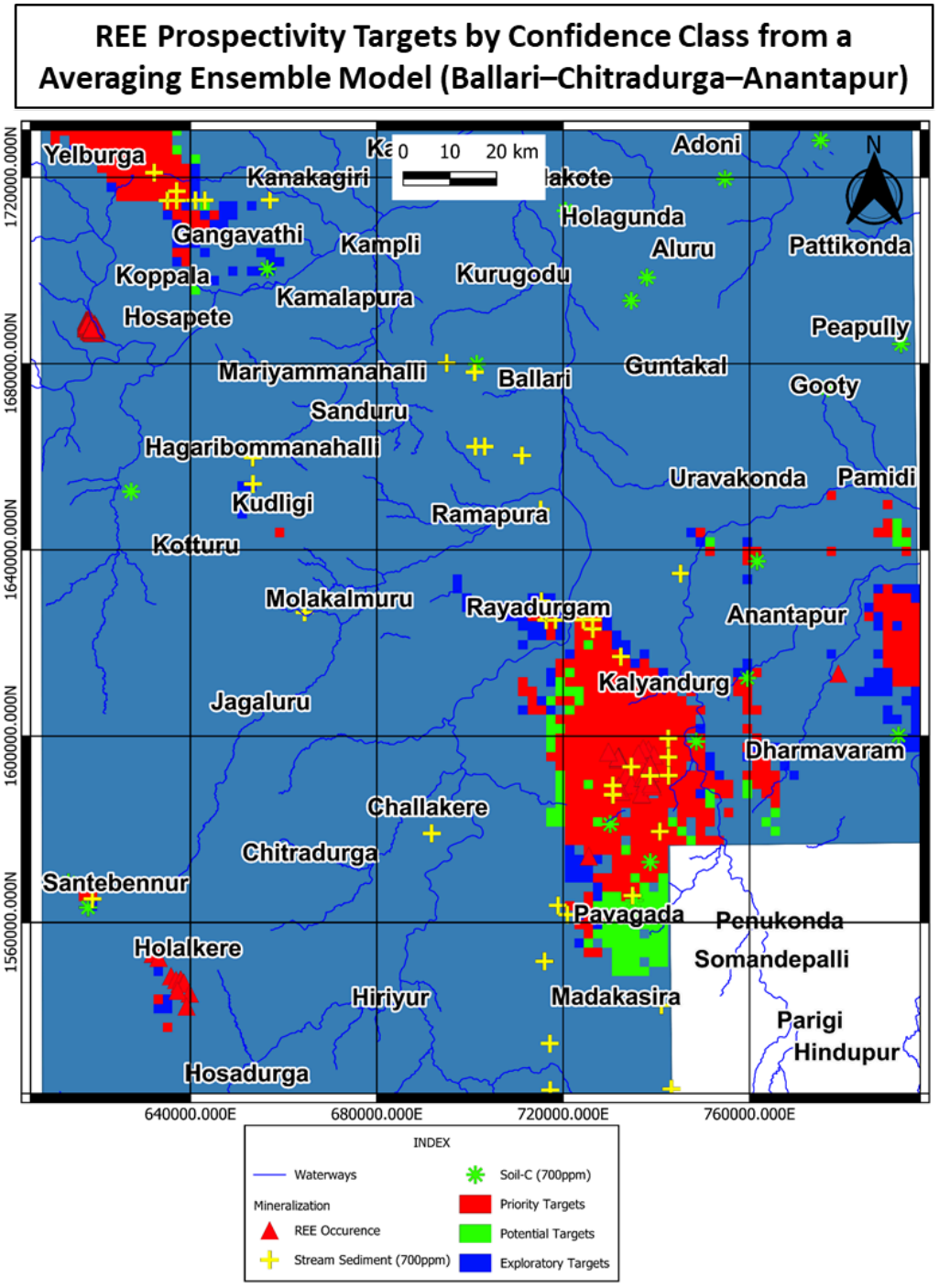
Bhukosh map displaying Confidence based REE Targets along Ballari-Chitradurga-Anatapur area, Karnataka & A.P



XIV. REE prospectivity targets for Stacked ensemble map



XV. REE prospectivity targets for Averaging ensemble map



XVI. Feature Importance Analysis

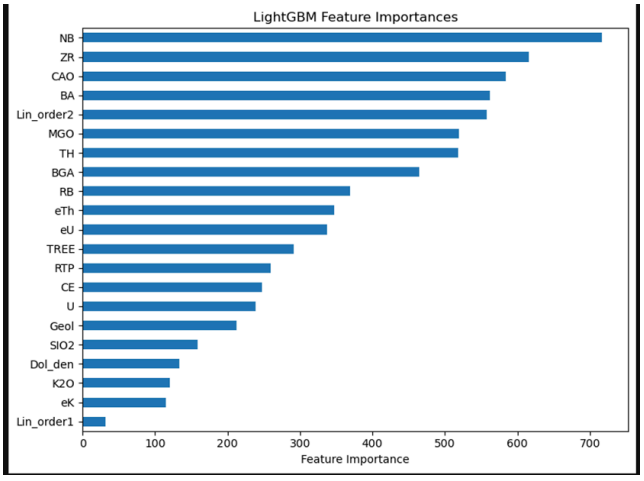
Understanding which predictors drive model decisions is crucial for both geological insight and practical targeting. Here, we compare feature-importance results from two contrasting training approaches:

1. **Soil-informed negative samples** (SS01_3k dataset) using a 3 km buffer around known REE sites.
2. **Random-negative sampling** (TS01_3k dataset) of background locations.

By examining the top contributors in each case—via LightGBM and XGBoost—we can both validate geochemical proxies for granite-hosted REE and assess how sampling strategy influences variable ranking.

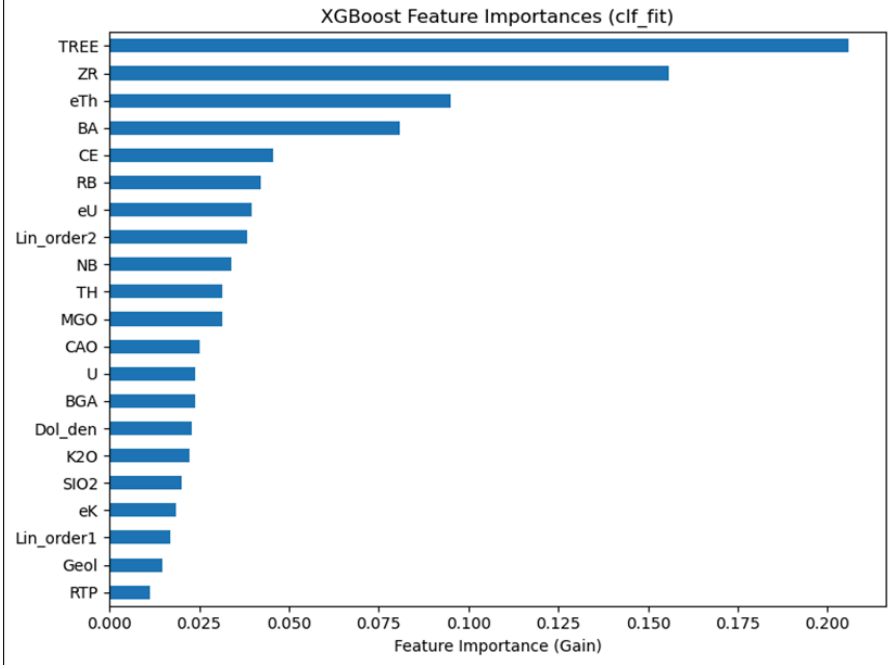
5.1 Feature Importance of Soil-Informed Models (SS01_3k)

LightGBM Top Five Features



1. **Nb (Niobium)**
   - High Nb is characteristic of evolved, peraluminous granites and pegmatites, marking REE-bearing intrusive centers.
2. **ZR (Zirconium)**
   - Zr concentrates in zircon during silica-rich magmatism; elevated Zr highlights felsic plutons.
3. **CAO (Calcium Oxide)**
   - Reflects plagioclase content and fractional crystallization trends often associated with late-stage REE enrichment.
4. **BA (Barium)**
   - Barium anomalies trace potassic alteration halos in K-feldspar–rich granitoids.
5. **Lin_order2 (Second-Order Lineament Density)**
   - Medium-scale structural corridors that localize hydrothermal fluids and REE-bearing vein emplacement.

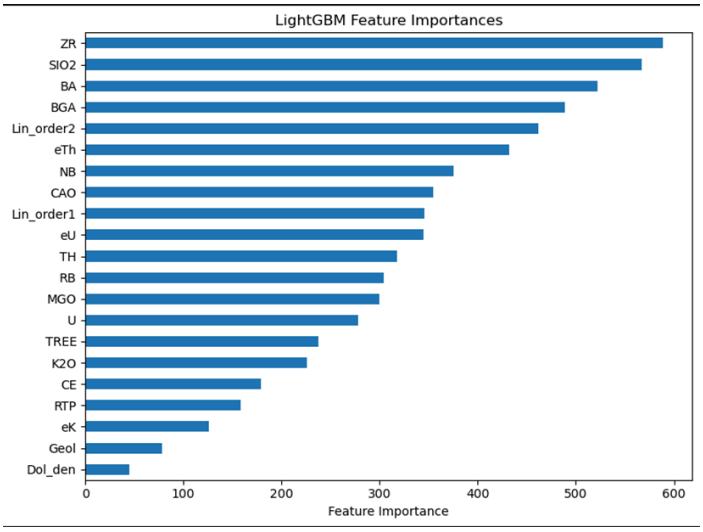
XGBoost Top Five Features



1. **TREE (Total REE content)**
    - Total REE content from Stream sediment data was used as direct proxy to target mineralization.
2. **ZR (Zirconium)**
    - Reinforces the zircon-rich signature of evolved granite bodies.
3. **eK (Potassium Gamma Counts)**
    - Direct proxy for K-feldspar abundance in felsic intrusions.
4. **BA (Barium)**
    - Again highlights potassic alteration and feldspar-rich zones favorable for REE.
5. **eTH (AGRS)**
    - Thorium substitutes into accessory minerals (e.g., monazite) in highly differentiated granites.

5.2 Feature Importance of Random-Negative Models (TS01_3k to TS01_5k)

LightGBM Top Five Features

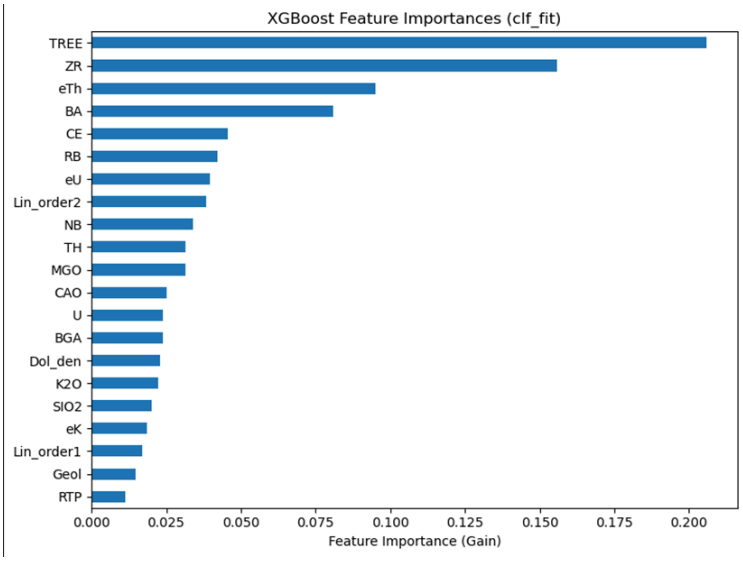


1. **ZR (Zirconium)**
    - Marks evolved felsic lithologies via zircon accumulation—strong granite tracer.

2. **SiO_2 (Silica)**
    - High SiO_2 identifies silica-saturated, highly differentiated granites often hosting REE.
3. **BA (Barium)**
    - Tracks K-feldspar/mica alteration halos around granitic intrusions.
4. **BGA (Residual Bouguer Anomaly)**
    - Gravity lows often coincide with less dense, highly fractured granite plutons.
5. **Lin_order2 (Second-Order Lineament Density)**
    - Medium-scale faults and joints that focus hydrothermal fluids and control emplacement.

> *Additional LightGBM contributors:* CaO, Nb, Lin_order1, TREE, RTP, eTh, Th.

XGBoost Top Five Features



1. ***TREE (Total REE content)**
    - Total REE content from Stream sediment data was used as direct proxy to target mineralization.
2. **BA (Barium)**
    - As above, a proxy for potassic alteration in granitoid bodies.
3. **eU (Potassium-corrected Uranium Counts)**
    - U anomalies flag accessory minerals (e.g., zircon, monazite) in evolved granites.
4. **eTh (Potassium-corrected Thorium Counts)**
    - Thorium enrichment in late-stage accessory phases reinforces granite delineation.
5. **Lin_order2 (Second-Order Lineament Density)**
    - Again highlights structural corridors for fluid flow and REE vein localization.

> *Additional XGBoost contributors:* Ce, Lin_order1.

XVII Recommendations

Future Work and Recommendations

The current study has demonstrated the utility of machine learning, including ensemble stacking approaches, for mineral prospectivity mapping (MPM) of rare earth elements (REE). However, several avenues remain for future improvement and refinement:

- **Expand Ground Truth Data**: The limited number of known mineralized sites restricts the training and validation of predictive models. Additional mineralized locations, identified through fieldwork and legacy data digitization, will significantly enhance model accuracy
- **Improve Sampling Resolution**: The existing geochemical datasets, particularly stream sediment and soil C-horizon samples, are based on a coarse sampling grid. This low spatial resolution likely limits their ability to detect subtle geochemical halos associated with REE mineralization. Notably, the Koppal region—despite confirmed mineralization—shows no significant REE signatures in the available stream sediment and soil data. This suggests a critical need for denser, high-resolution, and multi-depth sampling to improve anomaly detection.
- **Add Multisource Data Layers**: Incorporation of additional data types—such as hyperspectral imagery, radiometric data, and alteration mapping—may offer complementary insights that are not captured by conventional geophysical and geochemical datasets.
- **Expand Model Diversity**: While this study utilized a stacking ensemble of Random Forest, XGBoost, LightGBM, and Logistic Regression models, further improvement could be achieved by introducing a broader range of models. This includes spatially-explicit learners, deep learning frameworks, and hybrid geostatistical-machine learning methods to capture more complex geological relationships.
- **Incorporate Spatial Cross-Validation**: Traditional validation methods may overestimate model performance due to spatial autocorrelation. Future work should incorporate spatial cross-validation techniques to ensure more realistic assessments of model generalizability.
- **Quantify Model Uncertainty**: While ensemble averaging provided a basic measure of variability, future studies should explore more robust uncertainty quantification approaches. Bayesian models, probabilistic frameworks, or Monte Carlo simulations can help communicate the confidence of predictions more effectively, aiding exploration decisions.

XVIII Acknowledgement

We sincerely acknowledge the **Director**, Atomic Minerals Directorate for Exploration and Research (AMD), for granting permission and providing institutional support to undertake and participate in this study. The encouragement and facilitation provided by the Directorate were crucial for the successful execution of our work. We also extend our gratitude to our colleagues and mentors at AMD for their valuable discussions, guidance, and technical support throughout the course of this study.