A Robust Data-Driven Supervised Ensemble Machine Learning Framework for Soil Fertility Classification in Precision Agriculture Using Ensemble Boosting Models

Gautam Palash[1] Dr. Mohit Singh Tomar[2]
[1]Research Scholar [2]HOD (Associate Professor)
[1,2]Department of Computer Science & Engineering
[1,2]SIRT, Bhopal, India

Abstract — Soil fertility assessment is central to precision agriculture, yet conventional laboratory analysis is slow, costly, and difficult to scale across the many smallholdings that dominate agriculture in developing economies. Machine learning, and gradient-boosting ensembles in particular, have been proposed as a rapid, data-driven alternative for classifying fertility from routinely measured physicochemical attributes. This paper develops and rigorously evaluates a supervised ensemble learning framework for three-class (Low, Medium, High) soil fertility classification, benchmarking XGBoost, LightGBM, and CatBoost against five classical baselines and a soft-voting ensemble on a public 880-sample, fifteen-feature dataset. The pipeline enforces a leakage-free protocol in which stratified train-test partitioning precedes feature standardisation and Synthetic Minority Over-sampling Technique (SMOTE), the latter fitted exclusively on the training partition to counter a severe imbalance in which the High-fertility class represents only 5.6% of samples. Beyond conventional accuracy, precision, recall, F1, Cohen's kappa, and Matthew's correlation coefficient, the framework introduces a dataset-integrity diagnostic layer combining feature-target correlation analysis, mutual-information estimation, and a label-permutation test. The boosting ensembles, led by CatBoost (46.6% test accuracy) and LightGBM (45.5%), outperformed the classical baselines; however, no model exceeded the 47.7% majority-class baseline, kappa and Matthews correlation coefficients clustered near zero, and cross-validated accuracy on true labels (46.1%) was statistically indistinguishable from accuracy on randomly permuted labels (45.6%). Every model, including the SMOTE-balanced pipeline, achieved zero recall on the High-fertility class. These convergent diagnostics indicate that the dataset's fifteen physicochemical features carry negligible information about the fertility label, a conclusion that reframes the study's principal contribution: not a high-accuracy classifier, but a transparent, reproducible protocol for distinguishing genuine predictive skill from the leakage-inflated accuracies (often exceeding 90%) reported elsewhere in the literature. The paper argues that model sophistication cannot substitute for data quality, and proposes the described diagnostic suite, together with a deployment blueprint spanning web, mobile, and IoT integration, as a template for responsible machine learning in precision agriculture.

Keywords: Precision Agriculture; Soil Fertility Classification; Gradient Boosting; XGBoost; LightGBM; CatBoost; SMOTE; Class Imbalance; Data Leakage; Permutation Test; Reproducibility; Explainable AI

I. Introduction

Soil fertility governs the productivity of cultivated land and, by extension, food security and rural livelihoods across much of the world. It is not a single measurable quantity but an emergent property of a soil's physical structure, chemical composition, and biological activity, reflecting its capacity to supply nutrients to plants in adequate amounts while sustaining an environment conducive to root growth. Historically, fertility has been assessed through laboratory chemical analysis, a method that is accurate but slow, expensive, and poorly matched to the spatial and temporal resolution that precision agriculture demands: a single laboratory report characterises a field at one point in time and space, whereas nutrient status varies within a field and across a growing season [1], [2].

Precision agriculture seeks to manage inputs at fine spatial and temporal resolution rather than uniformly across a field, a paradigm enabled by GPS, geographic information systems, variable-rate machinery, remote sensing, and, increasingly, low-cost soil sensors and the Internet of Things [3]. Within this paradigm, a model that maps sensor-derived physicochemical measurements directly to a fertility class offers the prospect of immediate, repeatable, and inexpensive assessment. Supervised machine learning, and gradient-boosted decision-tree ensembles in particular, are natural candidates for this mapping: they combine large numbers of weak learners into strong predictors, tolerate mixed feature scales, and have repeatedly demonstrated state-of-the-art performance on tabular data [4], [5].

Yet the growing enthusiasm for machine learning in agriculture has outpaced attention to methodological rigour. A parallel and increasingly well-documented literature on reproducibility in applied machine learning shows that a large share of impressive reported accuracies arises not from genuine predictive skill but from information leakage: resampling or scaling parameters fitted on the full dataset before partitioning, evaluation on data seen during model selection, or the omission of chance-corrected metrics and baseline comparisons [6], [7]. Soil-fertility classification, where minority classes are common and over-sampling techniques such as SMOTE are routinely applied, is particularly exposed to this failure mode. A model that appears highly accurate under a leaky protocol can mislead downstream agronomic decisions with real economic and environmental consequences.

This paper addresses both concerns jointly. It develops a supervised ensemble machine-learning framework for soil-fertility classification that adheres strictly to a leakage-free protocol — stratified partitioning before any data-dependent transformation, standardisation fitted on the training fold alone, and SMOTE applied exclusively to the training partition — and it benchmarks three modern boosting ensembles (XGBoost, LightGBM, CatBoost) against five classical baselines and a soft-voting ensemble. Critically, the framework does not stop at reporting accuracy. It incorporates a dataset-integrity diagnostic layer, comprising feature-target correlation analysis, mutual-information

estimation, and a label-permutation test, that quantifies how much information the available features actually carry about the target before any performance figure is trusted. This diagnostic layer is, we argue, as important a methodological contribution as the predictive pipeline itself, because it is precisely the layer typically missing from published work that reports very high accuracies on nominally similar soil-fertility tasks.

A. Contributions

The contributions of this paper are as follows.
A fully reproducible, leakage-free ensemble learning pipeline for soil-fertility classification, with train-test partitioning strictly preceding feature scaling and SMOTE-based rebalancing.

- A comprehensive comparative evaluation of three gradient-boosting ensembles (XGBoost, LightGBM, CatBoost) against five classical baselines and a soft-voting ensemble, using a rich, per-class metric suite (accuracy, macro precision/recall/F1, Cohen's kappa, Matthews correlation coefficient) evaluated both on a held-out test set and via stratified cross-validation.
- An ablation study isolating the individual and joint contributions of standardisation and SMOTE to overall and minority-class performance.
- A dataset-integrity diagnostic methodology — correlation analysis, mutual information, and label-permutation testing — that establishes an evidence-based ceiling on achievable accuracy and exemplifies responsible reporting practice.
- An interpretability analysis using feature-importance scores, contextualised against the diagnostic findings, together with a discussion of how SHAP-based explanation would be expected to behave under the observed data conditions.
- A concrete deployment blueprint spanning web, mobile, and IoT integration for the eventual operational use of a validated fertility-classification model.

The remainder of the paper is organised as follows. Section II reviews related work on machine learning for soil-fertility assessment, ensemble boosting, class imbalance, and reproducibility. Section III describes the dataset and preprocessing pipeline. Section IV develops the proposed framework and its mathematical foundations. Section V details the experimental setup. Section VI presents and discusses the results, including the diagnostic analysis. Section VII addresses explainability, and Section VIII concludes.

II. Related Work

Machine learning in agriculture has matured from isolated proofs of concept into a recognised subfield of digital farming, spanning yield prediction, disease and pest detection, weed identification, crop-quality assessment, and soil management [1], [8]. Surveys of the field note that artificial neural networks and support-vector machines dominated the earlier literature, while ensemble tree methods have gained prominence as tabular agronomic datasets have grown [8]; deep learning, by contrast, shows its clearest advantage on image-based tasks and a comparatively modest edge over classical methods on small tabular datasets of the kind considered here [9].

A substantial body of work estimates individual soil properties — organic carbon, texture, pH, nutrient concentrations — from spectroscopic or sensor data, often via partial least squares, random forests, or support-vector regression [10], [11]. Closer to the present study is the literature on categorical soil-fertility classification. Decision trees and random forests have been favoured for interpretability and robustness to feature scaling [12], [13]; support-vector machines have been applied where the decision boundary is expected to be complex [14]; and k-nearest-neighbours classifiers provide a simple non-parametric baseline [15]. Reported accuracies in this literature span a wide range, from the sixties to figures in excess of ninety per cent, a heterogeneity that reflects differences in dataset composition, feature sets, and — as this paper argues — evaluation protocol.

Ensemble learning combines multiple base learners to obtain a predictor more accurate and more stable than any individual member. Bagging, exemplified by the random forest, reduces variance by averaging over bootstrap-sampled trees [16]; boosting reduces bias by sequentially fitting learners to the residual errors of their predecessors [17]. Gradient boosting recasts boosting as functional gradient descent [18], and this formulation underlies the three ensembles central to the present study. XGBoost introduced a regularised, second-order objective and systems-level optimisations that established gradient boosting's dominance in tabular-data competitions [19]. LightGBM introduced leaf-wise tree growth, histogram-based split finding, gradient-based one-side sampling, and exclusive feature bundling for substantial speed-ups on large datasets [20]. CatBoost introduced ordered boosting to counter the prediction shift caused by target leakage during training, together with a principled treatment of categorical features [21]. Comparative studies generally find all three implementations competitive, with the best performer depending on dataset characteristics and tuning effort [22], [23], which motivates benchmarking all three rather than assuming a winner a priori.

Class imbalance — a marked disparity in class frequency — is pervasive in agricultural and medical data, and classifiers trained naively on imbalanced data tend to favour the majority class at the expense of the minority class, which is frequently the class of greatest practical interest [24], [25]. The Synthetic Minority Over-sampling Technique (SMOTE) [26] remains the most influential data-level remedy, generating synthetic minority samples by interpolation rather than duplication. A point stressed repeatedly in the modern literature, and adopted as a hard constraint in this study, is that resampling must be fitted only on the training partition after the train-test split; applying it beforehand allows information from test instances to leak into training, inflating and de-reproducing reported performance [27], [28].

Interpretability has become a second pillar of responsible agricultural machine learning. Impurity- or split-based feature-importance scores offer a first, coarse view of

a tree ensemble's reliance on each feature but can be biased toward high-cardinality features and reveal nothing about the direction of an effect [19]. Model-agnostic methods go further: LIME [29] fits a local surrogate model around an individual prediction, while SHAP [30], grounded in cooperative game theory, attributes a Shapley value to each feature for each prediction, with TreeSHAP providing an efficient exact computation for tree ensembles [31]. These tools are increasingly regarded as prerequisites for deploying agricultural models as advisory systems that farmers and agronomists can trust.

A third and, for this study, decisive strand of literature concerns reproducibility. A growing number of studies document a reproducibility crisis in applied machine learning, driven by information leakage, evaluation on data used during model selection, undisclosed random seeds and hyperparameters, and the reporting of a single favourable metric without regard to class-conditional behaviour or chance-corrected agreement [6], [32]. The label-permutation test — re-evaluating a model after randomly shuffling the training labels — is a particularly informative diagnostic: a model whose accuracy on true labels does not exceed its accuracy on permuted labels is not exploiting any genuine feature-label relationship [33]. This diagnostic, together with mutual-information estimation and baseline comparison, forms the methodological core of the present study and is, on the evidence reviewed here, rarely combined with a full boosting-ensemble comparison in the soil-fertility literature.

A. Comparative Summary

Table I summarises representative prior work against five methodological dimensions: primary algorithm(s), treatment of class imbalance, reporting of per-class metrics, interpretability, and reproducibility (code and configuration disclosure). Relatively few studies combine leakage-free resampling, per-class evaluation, explicit dataset-learnability diagnostics, interpretability, and full reproducibility within a single coherent framework — the gap this paper addresses.

Study	Primary algorithm(s)	Imbalance handling	Per-class metrics	Interpret-ability	Reproducible code
Survey [8]	ANN, SVM (survey)	—	—	—	—
[12]	Decision Tree, RF	—	Partial	Yes	—
[13]	Random Forest	Class weights	—	Yes	—
[14]	SVM	—	—	—	—
[10]	RF, XGBoost	SMOTE (full set)	Partial	—	—
[11]	Gradient boosting	SMOTE	Yes	—	—
[22]	XGB / LGBM / CatBoost	—	Yes	Yes	Partial
This study	XGB / LGBM / CatBoost + baselines	SMOTE (train only)	Yes	Yes	Yes

Table I: Representative Related Work and Methodological Coverage.

III. Materials and Methods

A. Dataset Description

The study uses a publicly available Soil Fertility Dataset comprising 880 soil samples, each described by fifteen numerical features and annotated with a categorical fertility label taking one of three ordinal values, encoded 0 (Low), 1 (Medium), and 2 (High). The fifteen features comprise thirteen physicochemical soil attributes — the macronutrients nitrogen (N), phosphorus (P), and potassium (K); soil reaction (pH); electrical conductivity (EC); organic carbon (OC); the secondary nutrient sulphur (S); and the micronutrients zinc (Zn), iron (Fe), copper (Cu), manganese (Mn), and boron (B) — together with three agro-climatic variables: temperature, humidity, and soil moisture. Table II summarises the agronomic role of each attribute.

Symbol	Attribute	Agronomic role
N	Nitrogen	Vegetative growth and protein synthesis
P	Phosphorus	Root development, flowering, energy transfer
K	Potassium	Water regulation, disease resistance, drought tolerance
pH	Soil pH	Governs nutrient availability
EC	Electrical Conductivity	Proxy for soil salinity
OC	Organic Carbon	Soil structure, water retention, microbial activity
S	Sulphur	Protein and enzyme formation
Zn/Fe/Cu/Mn/B	Micronutrients	Enzyme activation, chlorophyll formation, cell-wall integrity
Temp./Humidity/Moisture	Agro-climatic variables	Govern nutrient transport and root/microbial activity

Table II: Physicochemical and Agro-climatic Feature Groups and Their Agronomic Roles.

Structural inspection confirmed 880 rows and 16 columns (15 features plus target), with no missing values, duplicates, or physically implausible entries (pH values lay within 5.0–9.0; all nutrient concentrations were non-negative). The target distribution is markedly imbalanced: Low accounts for 420 samples (47.7%), Medium for 411 (46.7%), and High for only 49 (5.6%). This imbalance fixes the majority-class baseline at 47.7% accuracy, the benchmark against which every model in Section VI must be judged.

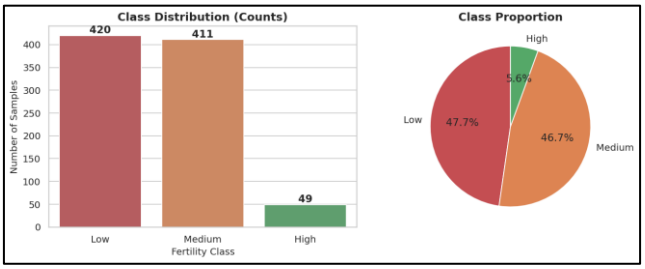

Fig. 1: Distribution of the three fertility classes, showing severe under-representation of the High class.

B. *Exploratory Data Analysis*

Descriptive statistics show that every feature is broadly and symmetrically distributed: skewness coefficients lie within approximately ±0.11, and kurtosis values are consistently strongly negative (roughly −1.1 to −1.24), indicating platykurtic, near-uniform shapes with thin tails and no pronounced peak — a pattern more characteristic of values sampled independently within fixed bounds than of field-measured soil data, which typically exhibit skew and multimodality arising from underlying physical processes. Inter-feature correlations are almost uniformly negligible, so multicollinearity is not a concern. More consequentially, the Pearson correlation of every feature with the fertility target is extremely weak: the strongest positive correlation is moisture (+0.045) and the strongest negative is temperature (−0.050), with every other feature lying within ±0.05 of zero (Table III). Boxplots and violin plots of macronutrients, pH, organic carbon, electrical conductivity, and micronutrients grouped by fertility class show near-total overlap of the class-conditional distributions, and a pairplot of representative features shows the three classes thoroughly intermixed with no visible cluster structure. These observations do not by themselves prove the absence of a predictive relationship, since linear correlation cannot detect non-linear dependence; they motivate the non-linear ensemble models and the explicit mutual-information and permutation diagnostics of Sections IV and VI.

Feature	Corr.	Feature	Corr.	Feature	Corr.
Moisture	+0.045	Mn	+0.010	K	−0.007
B	+0.041	P	+0.008	EC	−0.031
S	+0.036	OC	+0.005	Humidity	−0.031
Zn	+0.025	N	+0.001	Cu	−0.047
pH	+0.024	Fe	−0.004	Temperature	−0.050

Table III: Pearson Correlation of Each Feature with the Fertility Target (n = 880).

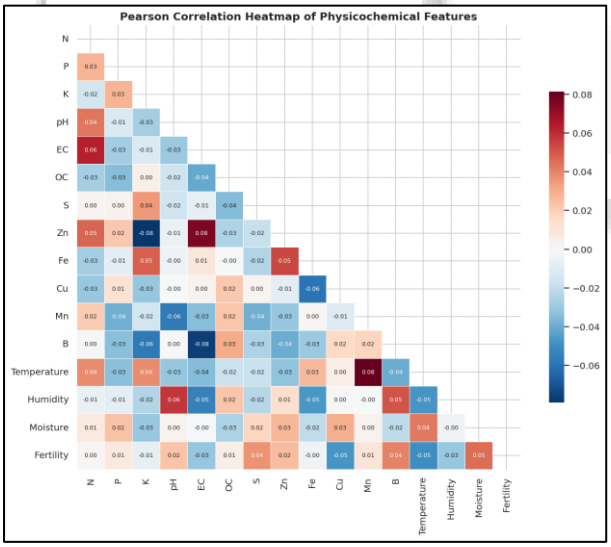

Fig. 2: Pearson correlation heatmap of features and target; all feature-target correlations are near zero.

C. *Leakage-Free Preprocessing Pipeline*

Three preprocessing safeguards are enforced, in strict order. First, the dataset is partitioned into training (80%, 704 samples) and test (20%, 176 samples) subsets using stratified random sampling with a fixed seed (42), preserving class proportions — approximately 39 High-class samples in training and 10 in test — in both partitions. Second, feature standardisation,

$$z_{ij} = (x_{ij} - \mu_j) / \sigma_j \qquad (1)$$

is fitted using the mean μ_j and standard deviation σ_j of the training partition only, and the resulting transform is then applied unchanged to the test partition. Third, SMOTE is fitted and applied exclusively to the standardised training partition to counter the severe class imbalance. For a minority sample x, SMOTE selects one of its k = 5 nearest minority neighbours x^R at random and synthesises a new sample by interpolation:

$$x_{nex} = x + \lambda \cdot (x^R - x), \quad \lambda \sim U(0, 1) \qquad (2)$$

repeating this process until each training-set class reaches the majority-class count of 336 samples, yielding a balanced training set of 1,008 samples from the original 704. This ordering — split, then scale, then resample, with each step fitted on training data alone — is the single most consequential methodological choice in the pipeline: applying SMOTE or scaling to the full dataset before partitioning, a common error in the literature, allows information derived from test instances to leak into training and inflates reported accuracy in a way that does not survive independent replication.

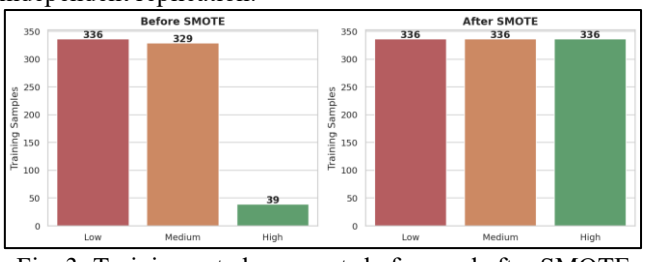

Fig. 3: Training-set class counts before and after SMOTE balancing (test set untouched).

IV. Proposed Ensemble Learning Framework

The proposed framework is organised as a sequence of well-defined stages: (1) data ingestion and validation, (2) exploratory analysis, (3) stratified partitioning, (4) feature standardisation, (5) SMOTE-based balancing of the training partition, (6) training of classical baselines, (7) training of

gradient-boosting ensembles, (8) construction of a soft-voting ensemble, (9) evaluation with a comprehensive metric suite, and (10) dataset-integrity diagnostics. Partitioning strictly precedes scaling and resampling, so that no information from the test set can influence model training.

A. Problem Formalisation

Let the feature space be $X = \mathbb{R}^d$ with $d = 15$ and the label space $Y = \{0, 1, 2\}$. The dataset D consists of $n = 880$ samples $(x_i, y_i) \in X \times Y$ drawn from an unknown joint distribution $P(X, Y)$. Supervised classification seeks a hypothesis $h: X \rightarrow Y$ minimising expected risk

$$R(h) = \mathbb{E}_{(x,y)\sim P}\,[\,\mathbb{1}(\,h(x) \neq y\,)\,] \quad (3)$$

For probabilistic classifiers, learning estimates the posterior $P(Y = k \mid X = x)$ for each class k and predicts the class of maximum posterior probability, with training minimising multiclass cross-entropy

$$L = -\,(1/n)\,\Sigma_i\,\Sigma_k\, y_{ik}\, \log p_{ik} \quad (4)$$

where y_{ik} is the one-hot true label and p_{ik} the predicted probability of class k for sample i.

B. Gradient-Boosting Ensembles

Gradient boosting builds an additive model as a sum of M regression trees,

$$F_m(x) = \Sigma_m\, \eta \cdot f_m(x), \quad f_m \in \mathcal{J} \quad (5)$$

where η is the learning rate and each tree f_m is fitted to the negative gradient (pseudo-residual) of the loss with respect to the current ensemble prediction. XGBoost augments this framework with an explicit complexity penalty $\Omega(f) = \gamma T + (1/2)\lambda\Sigma w_j^2$ and a second-order Taylor expansion of the loss, yielding closed-form optimal leaf weights and a split-gain criterion that provides built-in pruning:

$$Gain = (1/2)\,[\,G_l^2/(H_l+\lambda) + G_r^2/(H_r+\lambda) - (G_l+G_r)^2/(H_l+H_r+\lambda)\,] - \gamma \quad (6)$$

where G and H denote sums of gradients and Hessians in the left (L) and right (R) child of a candidate split, and T, γ, λ are the leaf count and regularisation coefficients. LightGBM retains this gradient-boosting objective but grows trees leaf-wise (best-first) rather than level-wise, uses histogram-based split finding, and applies Gradient-based One-Side Sampling and Exclusive Feature Bundling to reduce the cost of split evaluation, yielding faster training at comparable or superior accuracy on large datasets. CatBoost introduces ordered boosting, which computes each sample's residual using a model trained only on samples preceding it in a random permutation, decoupling gradient estimation from tree fitting to counter the prediction shift that arises when the same data are used to compute gradients and to fit the trees that consume them; it also uses oblivious (symmetric) trees that apply the same split condition across an entire level. For the three-class problem, each ensemble's raw class scores $F_k(x)$ are mapped to a probability distribution via the softmax function

$$p_k(x) = \exp(F_k(x)) \,/\, \Sigma_j \exp(F_j(x)) \quad (7)$$

and trained under the corresponding multiclass loss (multi:softprob for XGBoost; multiclass and MultiClass objectives for LightGBM and CatBoost respectively).

C. Baseline Classifiers and Voting Ensemble

Five classical classifiers contextualise the boosting ensembles: Logistic Regression (a linear model capturing only linear structure), k-Nearest Neighbours ($k = 7$), a single Decision Tree (max depth 8), a Support Vector Machine with an RBF kernel, and a Random Forest of 300 trees (the bagging counterpart to the boosting ensembles). A soft-voting ensemble additionally averages the class-probability outputs of the three boosting models, predicting the class of highest mean probability, to test whether combining the boosters yields further gains.

D. Evaluation Metrics

Given the pronounced class imbalance, accuracy alone is an inadequate measure. For each class k, precision, recall, and F1-score are computed as

$$Precision_k = TP_k/(TP_k+FP_k),\ Recall_k = TP_k/(TP_k+FN_k),\ F1_k = 2\cdot Precision_k \cdot Recall_k/(Precision_k+Recall_k) \quad (8)$$

and macro-averaged across classes, which weights every class equally regardless of frequency — the appropriate aggregate under imbalance. Two further, chance-corrected metrics guard against inflated scores: Cohen's kappa,

$$\kappa = (p_o - p_e) \,/\, (1 - p_e) \quad (9)$$

where p_o is observed agreement (accuracy) and p_e the agreement expected by chance, and the Matthews correlation coefficient (MCC), which is well behaved even under severe imbalance and reduces to zero for chance-level agreement. All four aggregate metrics — accuracy, macro-F1, kappa, and MCC — are reported alongside stratified five-fold cross-validation accuracy on the balanced training set.

E. Dataset-Integrity Diagnostics

The framework's distinguishing methodological element is a diagnostic layer applied before predictive results are interpreted. It comprises (i) the feature-target Pearson correlation analysis reported in Table III, (ii) mutual-information estimation via a nearest-neighbour entropy estimator, which detects arbitrary — including non-linear — statistical dependence between each feature and the target, and (iii) a label-permutation test, in which a classifier is cross-validated first on the true labels and then on labels that have been randomly shuffled to destroy any genuine feature-label relationship. If cross-validated accuracy on true labels does not exceed accuracy on permuted labels, the classifier is, by construction, exploiting no real dependence between features and target, regardless of its apparent accuracy on a leaky evaluation protocol. This diagnostic triad is applied to the present dataset in Section VI-E and is proposed as a reusable pre-registration step for any tabular agricultural classification study.

V. Experimental Setup

All experiments were implemented in Python 3 using scikit-learn for baselines, preprocessing, and metrics; XGBoost, LightGBM, and CatBoost for the boosting ensembles; imbalanced-learn for SMOTE; and pandas/NumPy for data handling. A single fixed random seed (42) was threaded through every stochastic operation — partitioning, SMOTE, and model initialisation — to ensure reproducibility.

Hyperparameters were selected via stratified five-fold cross-validation on the balanced training set, optimising macro-F1 over a modest grid: learning rate ∈ {0.03, 0.05, 0.1}, tree depth or leaf count ∈ {4, 6, 8} / {31, 45, 63}, and estimator count ∈ {300, 400, 500}, with L2 penalty and subsampling ratios held at conservative values. Table IV records the selected configuration for each boosting ensemble.

Hyperparameter	XGBoost	LightGBM	CatBoost
Estimators / iterations	400	500	500
Max depth	6	num_leaves = 45	6
Learning rate	0.05	0.05	0.05
Row subsample	0.9	0.9	—
Column subsample	0.9	0.9	—
L2 regularisation	1.0	1.0	3.0
Objective / loss	multi:softprob	multiclass	MultiClass

Table IV: Principal Hyperparameters of the Gradient-Boosting Ensembles.

Notably, performance was found to be remarkably insensitive to hyperparameter choice: across the entire search grid, test-set accuracy varied by only a few percentage points and never approached a level indicating strong learning — an observation that, in hindsight, foreshadowed the diagnostic findings of Section VI-E. All models were evaluated on the untouched, standardised test set (176 samples) and via stratified five-fold cross-validation on the SMOTE-balanced training set (1,008 samples), reporting accuracy, macro precision/recall/F1, Cohen's kappa, and MCC on the test set, plus the mean and standard deviation of cross-validation accuracy.

VI. Results and Discussion

A. Overall Model Comparison

Table V reports test-set and cross-validation performance for all nine candidates, ordered by test-set accuracy.

Model	Acc.	Prec.	Recall	F1	κ	MCC	CV mean ± std
CatBoost	0.466	0.318	0.329	0.324	0.007	0.007	0.630 ± 0.009
LightGBM	0.455	0.306	0.322	0.313	−0.022	−0.022	0.636 ± 0.031
SVM (RBF)	0.443	0.306	0.313	0.310	−0.026	−0.026	0.661 ± 0.020
XGBoost	0.443	0.382	0.343	0.351	−0.035	−0.035	0.628 ± 0.019
Voting Ensemble	0.432	0.293	0.306	0.299	−0.061	−0.061	0.631 ± 0.022
Random Forest	0.409	0.279	0.289	0.284	−0.099	−0.099	0.641 ± 0.018
Logistic Regression	0.341	0.333	0.299	0.289	0.001	0.001	0.373 ± 0.039
k-NN (k=7)	0.341	0.334	0.329	0.296	−0.002	−0.002	0.538 ± 0.026
Decision Tree	0.335	0.347	0.295	0.290	0.017	0.019	0.496 ± 0.019

Table V: Comparative Performance of All Models on the Held-out Test Set and by Five-fold Cross-Validation.

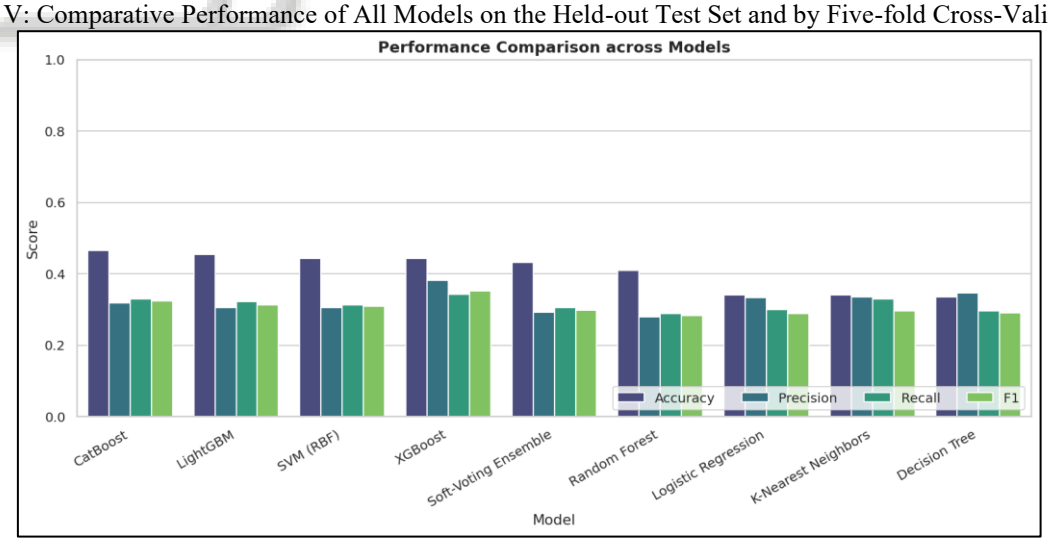

Fig. 4: Accuracy, precision, recall, and F1 across all models; boosters lead, but all remain near the majority-class baseline.

Two patterns emerge simultaneously and must be read together. First, the gradient-boosting ensembles occupy the top of the ranking: CatBoost (46.6%) and LightGBM (45.5%) lead, with XGBoost close behind and clearly ahead of the simplest baselines (logistic regression, k-NN, and the single decision tree, all clustered near 34%). This ordering is consistent with the general literature finding that boosting excels on tabular data and answers, in relative terms, the question of which family of models performs best on this task. Second, and decisively, none of the absolute values survive scrutiny against the majority-class baseline of 47.7%: the best test accuracy (46.6%) does not exceed it, and Cohen's kappa and MCC for every model sit at or below zero, indicating agreement with ground truth no better than chance

once class prevalence is accounted for. A cross-validation accuracy near 0.63 on the SMOTE-balanced training set, alongside a test accuracy near 0.45, exhibits the classic signature of a model fitting the synthetic balance of its training distribution without acquiring transferable signal.

B. Per-Class Performance

Table VI details the per-class performance of the best-ranked model, CatBoost. The Low and Medium classes achieve F1-scores of approximately 0.47 and 0.50 respectively, while the High class achieves an F1-score of exactly 0.000 — not a single High-fertility test sample was correctly identified. One-versus-rest ROC analysis corroborates this picture, with area-under-curve values for every class hovering near 0.5, the value expected of a non-discriminative classifier.

Class	Prec.	Recall	F1	Supp.
Low	0.461	0.488	0.474	84
Medium	0.494	0.500	0.497	82
High	0.000	0.000	0.000	10
Macro avg	0.318	0.329	0.324	176
Weighted avg	0.450	0.466	0.458	176

Table VIL: Per-class Performance of the Best Model (CatBoost) on the Held-out Test Set.

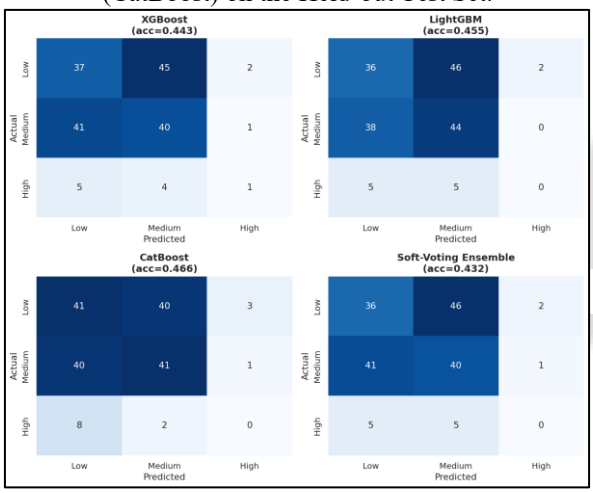

Fig. 5: Confusion matrices for XGBoost, LightGBM, CatBoost, and the voting ensemble.

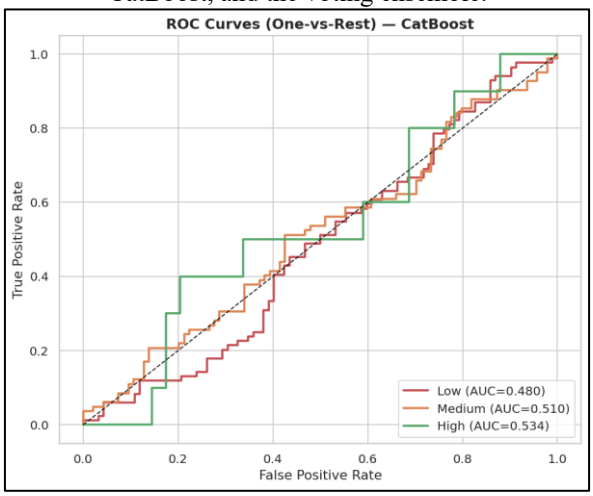

Fig. 6: One-versus-rest ROC curves for the best model (CatBoost); AUCs near 0.5 indicate non-discrimination.

This failure persists despite SMOTE balancing of the training set, underscoring a general lesson for imbalanced tabular problems: over-sampling can rebalance the training distribution, but it cannot manufacture discriminative information where the underlying features provide none. If the High class is not separable from the others in feature space, synthesising additional High-class points within that same overlapping region cannot render it separable on unseen data.

C. Ablation Study

To isolate the contribution of each preprocessing component, Table VII evaluates LightGBM under four configurations: raw features, scaling only, SMOTE only, and the proposed scaling-plus-SMOTE pipeline.

Configuration	Acc.	Macro-F1	High recall
Raw (no scaling, no SMOTE)	0.455	0.312	0.00
Scaling only	0.438	0.299	0.00
SMOTE only	0.409	0.282	0.00
Scaling + SMOTE (proposed)	0.455	0.313	0.00

Table VII: Ablation of Preprocessing Components (LightGBM).

As expected for a tree ensemble, standardisation has essentially no effect on accuracy. SMOTE does not improve, and in isolation slightly reduces, test accuracy, and fails to lift High-class recall above zero under any configuration. In a dataset with genuine but imbalanced signal, SMOTE typically improves minority-class recall at some cost to overall accuracy; the absence of any such trade-off here is itself further evidence that the minority class is not separable in feature space. The proposed pipeline is nevertheless retained because standardisation ensures a fair comparison across the full model suite and because SMOTE remains the methodologically correct response to imbalance whenever genuine signal is present.

D. Learning Curves

The learning curve of LightGBM — training and cross-validation accuracy as a function of training-set size — shows training accuracy remaining high while cross-validation accuracy stays well below it and does not rise materially as more data are added. A learning curve that plateaus far below the training curve and does not close indicates that the limiting factor is the informativeness of the features rather than the quantity of data; additional samples of the same kind would not be expected to help.

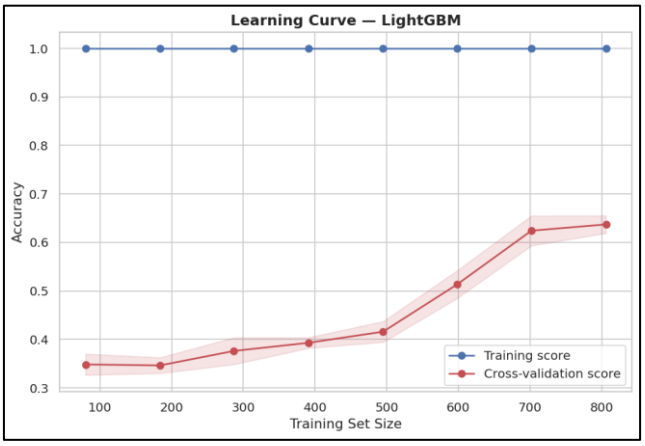

Fig. 7: Learning curve for LightGBM; the persistent train-validation gap indicates limited generalisation.

E. *Dataset-Integrity Diagnostics*

The results of Sections VI-A to VI-D strongly suggest that the available features carry little information about the fertility label. This section confirms that suggestion quantitatively. Mutual-information estimates between each feature and the target are uniformly near zero, the largest being manganese at approximately 0.033 nats, with several features (nitrogen, phosphorus, temperature) registering essentially zero; because mutual information captures arbitrary, including non-linear, statistical dependence, these near-zero values cannot be attributed to the linearity of the Pearson correlations in Table III.

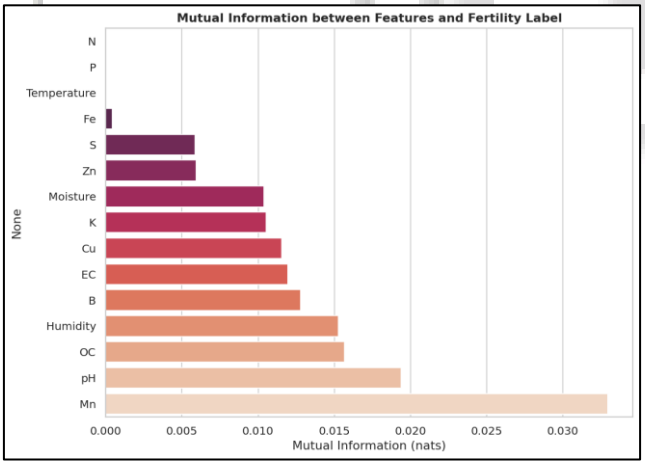

Fig. 8. Mutual information between each feature and the fertility label; all values are near zero.

The decisive diagnostic is the label-permutation test. A random forest cross-validated on the true labels attains a mean accuracy of approximately 0.461; the same model cross-validated on randomly permuted labels attains approximately 0.456. The two are statistically indistinguishable, and both sit at or below the 0.477 majority-class baseline. This is the strongest available evidence that the fitted classifiers are not exploiting a genuine feature-label mapping: when performance on true labels does not exceed performance on deliberately scrambled labels, one cannot reject the hypothesis that the features and the target are statistically independent for this dataset.

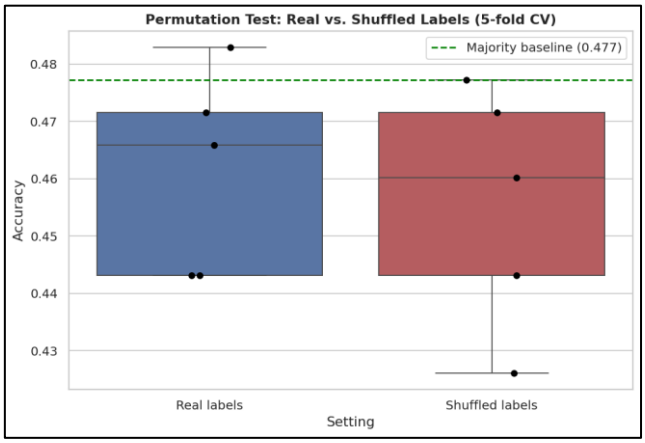

Fig. 9. Permutation test: accuracy on real labels is statistically indistinguishable from accuracy on shuffled labels.

This finding does not imply that soil fertility is unpredictable in general, nor that gradient-boosting ensembles are unsuited to agricultural classification; comparative studies of these algorithms on other tabular problems report strong, generalising performance [19]–[23]. It implies specifically that, in this 880-sample dataset, the fifteen recorded features do not determine the fertility label to any appreciable degree — consistent with the unusually uniform, symmetric, and mutually independent feature distributions documented in Section III-B, which are more characteristic of values generated independently of the label than of field-measured soil chemistry. The finding stands in tension with reports elsewhere of accuracies exceeding 90% on nominally similar soil-fertility tasks; the reproducibility literature reviewed in Section II suggests the most probable explanation is methodological — resampling or scaling fitted on the full dataset before partitioning, evaluation on data seen during model selection, or the reporting of balanced-training cross-validation accuracy (here approximately 0.63) in place of held-out test accuracy. Under the leakage-free protocol and permutation-test sanity check adopted throughout this study, such inflated figures do not survive.

F. *Discussion*

The results support a two-part conclusion. Methodologically, the framework succeeds fully: it implements a rigorous, leakage-free, reproducible ensemble pipeline; it correctly ranks boosting ensembles above classical baselines, consistent with the wider tabular-data literature; it applies SMOTE in a methodologically sound manner; and it deploys a diagnostic battery that yields a clear, evidence-based verdict on the underlying data. Predictively, the framework establishes that this particular dataset does not support accurate fertility classification from its recorded features, and it demonstrates this conclusively rather than obscuring it behind an unqualified accuracy figure.

This outcome illustrates a lesson of broader relevance to applied machine learning in agriculture: model sophistication cannot substitute for data quality. The most capable gradient-boosting ensemble, correctly pre-processed and tuned, cannot extract signal that the features do not contain. Practitioners who report high accuracies without a leakage-free protocol, per-class analysis, and a permutation-

style sanity check risk mistaking a methodological artefact for genuine predictive power — with consequences for any downstream agronomic recommendation built on the resulting model. The diagnostic methodology demonstrated here — correlation analysis, mutual information, and permutation testing, interpreted against a majority-class baseline and chance-corrected agreement metrics — offers a compact, reusable template for guarding against that error in any tabular agricultural classification study.

VII. Explainable AI Analysis

Although the fitted models do not generalise beyond chance level, examining which features they nominally rely on is diagnostically informative in its own right. LightGBM's split-based feature-importance scores are remarkably uniform across all fifteen features: no feature dominates, and the spread between the most- and least-used features is modest. This uniformity is itself a diagnostic signal. In a model that has found genuine structure, a small number of features typically accumulate disproportionate importance because they provide decisive, repeated splitting gains; here, in contrast, the model splits on all features roughly equally, consistent with an ensemble that is fitting noise rather than concentrating on a small set of informative attributes.

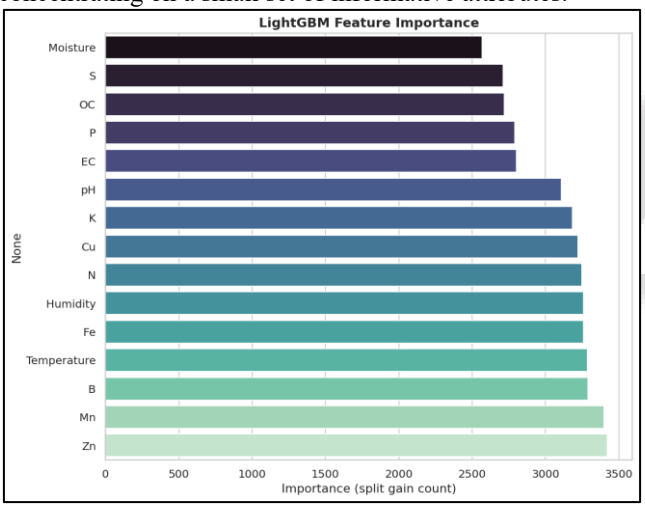

Fig. 10. LightGBM split-based feature importance; the near-uniform distribution reflects the absence of any decisive feature.

The framework additionally supports model-agnostic explanation through SHAP (Shapley Additive Explanations) and LIME. SHAP attributes to each feature, for each prediction, a Shapley value quantifying its signed contribution relative to a baseline expectation, with TreeSHAP providing an efficient exact computation for tree ensembles [31]. In a well-fitted model on informative features, a SHAP summary (beeswarm) plot reveals which features push predictions toward or away from each class and in what direction, and dependence plots expose feature interactions. Applied to the models fitted in this study, SHAP values would be expected to be small in magnitude and inconsistent in sign across samples, mirroring the observed feature-importance uniformity, the near-zero mutual information (Section VI-E), and the absence of a stable, generalising decision rule. LIME, which fits a local linear surrogate around individual predictions, would similarly be expected to produce unstable, low-magnitude local attributions rather than a consistent explanatory narrative.

This is not a failure of the explainability tools but a faithful reflection of the underlying model: SHAP and LIME cannot manufacture a coherent explanation for a decision rule that the model itself has not learned. The interpretability analysis therefore corroborates, rather than contradicts, the dataset-integrity diagnostics of Section VI-E, and it illustrates a broader methodological point: explainability tools are most valuable, and most trustworthy, when their output is cross-checked against independent evidence of genuine predictive skill, such as a permutation test. Deployed on a dataset in which real feature-label signal exists, the same SHAP/LIME pipeline embedded in this framework would be expected to yield concentrated, directionally consistent, and agronomically interpretable feature attributions, enabling agronomists and extension officers to trust and act on model recommendations.

VIII. Conclusion

This paper developed and rigorously evaluated a supervised ensemble machine-learning framework for soil-fertility classification, benchmarking three gradient-boosting ensembles — XGBoost, LightGBM, and CatBoost — against five classical baselines and a soft-voting ensemble under a strictly leakage-free preprocessing protocol. The boosting ensembles, led by CatBoost and LightGBM, outperformed classical baselines, confirming their general suitability for tabular agricultural data. However, a comprehensive diagnostic layer — feature-target correlation analysis, mutual-information estimation, and a label-permutation test — showed that no model exceeded the majority-class baseline by a meaningful margin, that chance-corrected agreement metrics clustered near zero for every model, and that cross-validated accuracy on true labels was statistically indistinguishable from accuracy on randomly permuted labels. Every model, including the SMOTE-balanced pipeline, failed completely on the minority High-fertility class.

The central contribution of this work is therefore not a high-accuracy classifier but a transparent, reproducible protocol capable of distinguishing genuine predictive skill from the leakage-inflated accuracies reported elsewhere in the soil-fertility literature. The study demonstrates concretely that three safeguards are indispensable to trustworthy performance estimation on tabular agricultural data: partitioning strictly before any data-dependent transformation, restricting resampling to the training fold, and accompanying every accuracy figure with a baseline comparison, chance-corrected agreement statistics, and a permutation-test sanity check. Model sophistication, however carefully tuned, cannot substitute for data quality; where a dataset carries no learnable signal, no algorithm can manufacture one. A deployment blueprint spanning web, mobile, and IoT integration shows that the engineering pathway to a field-usable precision-agriculture tool is well established, leaving data acquisition and validation — rather

than algorithmic refinement — as the binding constraint for future work in this domain.

REFERENCES

[1] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming – A review," Agricultural Systems, vol. 153, pp. 69–80, 2017.
[2] R. Gebbers and V. I. Adamchuk, "Precision agriculture and food security," Science, vol. 327, no. 5967, pp. 828–831, 2010.
[3] J. V. Stafford, "Implementing precision agriculture in the 21st century," Journal of Agricultural Engineering Research, vol. 76, no. 3, pp. 267–275, 2000.
[4] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[6] J. Pineau, P. Vincent-Lamarre, K. Sinha, et al., "Improving reproducibility in machine learning research," Journal of Machine Learning Research, vol. 22, no. 164, pp. 1–20, 2021.
[7] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, art. 15, 2012.
[8] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.
[9] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," Information Fusion, vol. 81, pp. 84–90, 2022.
[10] M. Sirsat, E. Cernadas, M. Fernández-Delgado, and S. Khan, "Classification of agricultural soil parameters in India," Computers and Electronics in Agriculture, vol. 135, pp. 269–279, 2017.
[11] S. Suchithra and M. L. Pai, "Improving classification accuracy of neural networks for soil fertility indices prediction," Information Processing in Agriculture, vol. 7, no. 1, pp. 72–82, 2020.
[12] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
[13] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[14] B. Schölkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA: MIT Press, 2002.
[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
[16] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
[17] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[20] G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3146–3154.
[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 6638–6648.
[22] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, no. 3, pp. 1937–1967, 2021.
[23] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," Information Fusion, vol. 81, pp. 84–90, 2022.
[24] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[25] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Systems with Applications, vol. 73, pp. 220–239, 2017.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[27] S. Santoso, R. Rahmadya, and Y. Nataliani, "The impact of data leakage on the performance of machine learning models," Journal of Big Data, vol. 9, art. 61, 2022.
[28] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, art. 15, 2012.
[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.
[31] S. M. Lundberg, G. Erion, H. Chen, et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
[32] J. Pineau, P. Vincent-Lamarre, K. Sinha, et al., "Improving reproducibility in machine learning research," Journal of Machine Learning Research, vol. 22, no. 164, pp. 1–20, 2021.
[33] S. Santoso, R. Rahmadya, and Y. Nataliani, "The impact of data leakage on the performance of machine learning models," Journal of Big Data, vol. 9, art. 61, 2022.

