Agronomy Research **24**(S2), xxx–ccc, 2026
https://doi.org/10.15159/AR.26.089

Predicting ginger rhizome disease severity from soil physicochemical parameters: A comparative study of core machine learning models

A.B. Sadavare[1,*] and P.P. Shinde[2]

[1]Department of Computer Science & Engineering, Shree Santkrupa Institute of Engineering & Technology, Ghogaon 415111, Maharashtra, India
[2]Department of Computer Applications (MCA), Government College of Engineering, Karad 415124, Maharashtra, India
[*]Correspondence: ashokskarad@gmail.com

Received: August 28th, 2026; Accepted: October 28th, 2026; Published: November 6th, 2026

Abstract. Rhizome soft rot is a major constraint on ginger (*Zingiber officinale*) production, yet machine learning studies on ginger health mostly use leaf images rather than soil conditions. This study tested whether routine soil measurements can predict rhizome disease severity, and identified a minimal set of measurements for field-deployable monitoring tools. Soil samples (n = 1,141) were collected from 901 plots in Satara district, Maharashtra, India, across three growth stages, and analysed for pH, moisture, temperature, nitrogen, phosphorus, potassium, organic matter, electrical conductivity (EC) and texture. Severity was scored as Healthy, Mild-Moderate or Severe, and pathogens were identified by molecular methods. Support Vector Machine, Random Forest, XGBoost, LightGBM and a soft Voting Ensemble were tuned and evaluated on a plot-wise test set (n = 231) and by repeated group-wise cross-validation. Test accuracy was 74.5–75.8% and macro F1-score 0.644–0.696, highest for XGBoost. Class weighting, SMOTE, ADASYN, Borderline-SMOTE and focal loss raised recall of the rare Severe class from about 0.32–0.39 to 0.47–0.57 but lowered precision, so macro F1-score did not improve. SHAP and LIME showed that high cumulative moisture, high EC and their product increased the probability of Severe disease. Spatial and temporal hold-out tests gave macro F1-scores of 0.64–0.70. Moisture and EC alone retained 84–90% of the full-model macro F1-score. Moisture and EC sensing therefore offers a basis for low-cost, field-deployable early warning of rhizome disease, pending validation in other regions and seasons.

Key words: class imbalance, disease severity classification, ensemble learning, feature importance, field-deployable monitoring, precision agriculture, rhizome rot, SHAP, soil electrical conductivity.

INTRODUCTION

Ginger (*Zingiber officinale* Roscoe) is a high-value rhizomatous crop cultivated widely across South and Southeast Asia for its culinary, medicinal and industrial uses. Because the crop is propagated vegetatively from seed rhizome rather than from true seed, the physiological quality and nutrient status of planting material, and the below-ground

conditions in which it develops, exert a disproportionate influence on final yield and marketable quality; optimisation of rhizome nutrition and propagation practice remains an active area of applied agronomic research (Wang et al., 2025a).

Soil-borne disease risk in rhizomatous and root crops is strongly conditioned by soil moisture regime and its electrochemical correlates. Soil electrical conductivity (EC) and dielectric behaviour respond directly to moisture content and fertiliser status, and are increasingly used as low-cost, sensor-derived proxies for below-ground conditions relevant to crop stress and predictive modelling (Palta et al., 2025). Waterlogged, poorly drained soils favour the motile, water-dependent zoospores of *Pythium* species, the causal agents of rhizome soft rot, so soil moisture status and its correlates are plausible early indicators of disease risk long before visible symptoms appear above ground.

Precision agriculture, underpinned by low-cost sensing and machine learning, offers a route to earlier and more objective assessment of below-ground crop stress than conventional field scouting. Artificial intelligence methods have been applied across a wide range of agricultural produce and production systems to enhance quality and yield outcomes (Nautiyal et al., 2025), to forecast crop performance and food security under drought-induced water stress (Fares et al., 2026), and, more fundamentally, to reconceive soil health as a multivariate, sensor-derived construct that machine learning can characterise more completely than a small panel of laboratory indices (Schweng et al., 2025).

Soil moisture and salinity, the two variables most directly implicated in waterlogging-associated disease risk, have been predicted successfully from remote sensing and in-situ sensor networks using a wide range of machine learning approaches. These include cloud-based mapping of soil salinity from satellite imagery (Raja Ranganathan et al., 2026), cascaded Internet-of-Things-based prediction of soil moisture and salinity in wetland systems (Song & Yi, 2025), and stacking ensemble models that fuse multi-source remote sensing data to improve salinity-inversion accuracy in salinised croplands (Deng et al., 2025). Related work has generated fine-resolution soil moisture products to support regional water management (Chen et al., 2025) and has coupled soil, environmental and crop-growth variables to reveal the spatiotemporal drivers of field-scale salinity (Xu et al., 2026).

Machine learning has similarly been used to estimate soil macronutrient status directly from spectroscopic and physicochemical data, reducing reliance on slow, destructive laboratory analysis (Huai et al., 2025), and to disentangle how land-use practices shape the physicochemical properties of agricultural soils more broadly (Ngu et al., 2025). Despite this substantial and rapidly growing body of soil-focused machine learning research, its application to ginger specifically has remained limited, and a soil-parameter-based approach to rhizome disease severity prediction, evaluated with algorithms robust to the class imbalance and correlated features typical of soil datasets, is, to our knowledge, unreported for this crop.

The present study addresses this gap. Using a soil physicochemical dataset linked to field-assessed and molecularly confirmed disease severity, we developed and comparatively evaluated five supervised classification algorithms – Support Vector Machine (SVM), Random Forest, XGBoost, LightGBM and a Voting Ensemble of the tree-based and kernel models – for three-class prediction of ginger rhizome disease severity. Model performance was assessed under both a conventional held-out test split and group-wise cross-validation that respected the repeated-measures structure of the data, and feature importance was examined across algorithms to identify which soil

parameters most reliably signal elevated disease risk. The overarching goal of the study is to identify a minimal set of soil variables that can support low-cost, field-deployable monitoring tools, as a precursor to resource-constrained (edge) implementation. Specifically, the study (i) benchmarks the five classifiers; (ii) tests whether class-imbalance handling (class weighting, SMOTE, ADASYN, Borderline-SMOTE and a focal-loss objective) improves detection of the rare Severe class; (iii) explains the predictions with SHAP and LIME; (iv) checks generalisation across space and growth stages; and (v) quantifies how few sensors are needed to retain useful accuracy.

MATERIALS AND METHODS

Overall framework

The study followed the workflow summarised in Fig. 1. Field soil data and disease scores were first compiled and converted into engineered features. The data were then partitioned by plot, the five classifiers were tuned and trained under several imbalance-handling strategies, and the resulting models were evaluated on a held-out test set, by repeated group-wise cross-validation and by spatial and temporal hold-out tests. Finally, the predictions were explained with SHAP and LIME, and nested feature subsets were used to identify a minimal sensor set suitable for field-deployable monitoring.

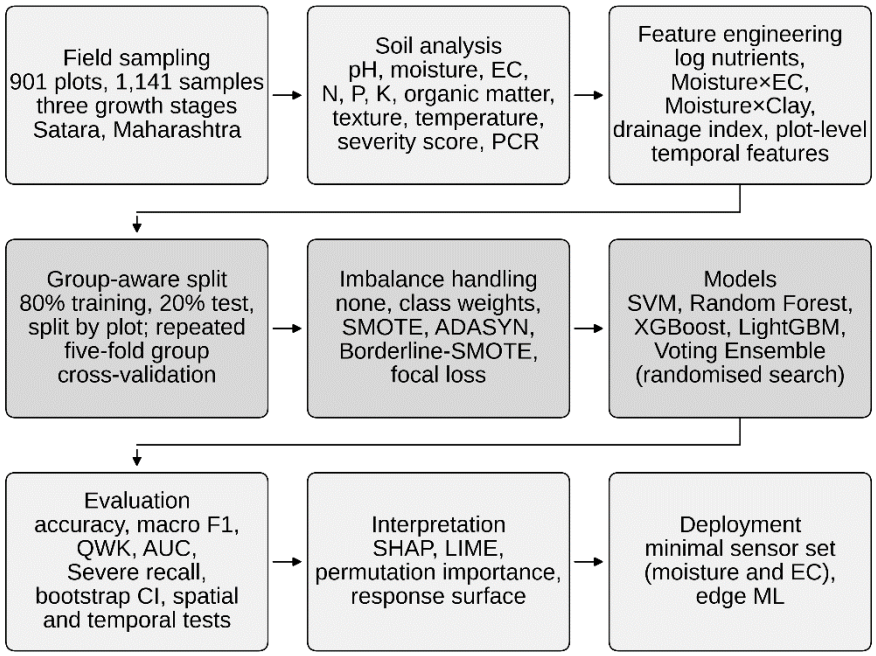

Figure 1. Overall framework of the study. Field samples (901 plots, 1,141 samples, three growth stages) were analysed in the laboratory and scored for disease severity. Engineered features fed five classifiers, which were tuned by randomised search with group-wise cross-validation and trained under six class-imbalance strategies. Models were evaluated on a plot-wise hold-out set, repeated group-wise cross-validation and spatial and temporal hold-out tests, interpreted with SHAP, LIME and permutation importance, and reduced to a minimal moisture-and-EC sensor set for field-deployable monitoring.

Study area and data collection

Soil and disease severity data were collected from ginger-growing plots in Satara district, Maharashtra, India (approximately 17.37–17.95° N), during a single cropping season. In total, 1,141 georeferenced soil samples were obtained from 901 distinct plots, with a subset of plots sampled repeatedly across three phenological stages: early vegetative growth, mid-season rhizome bulking, and pre-harvest (120 plots were sampled at all three stages and 781 plots once). At each sampling event, soil was characterised for pH, gravimetric moisture content (%), temperature (°C), plant-available nitrogen, phosphorus and potassium (mg kg^{-1}), organic matter content (%), electrical conductivity (EC, dS m^{-1}), and particle-size fractions (sand, silt and clay, %), from which a USDA textural class was assigned. Plant disease status was scored in the field on a three-class ordinal severity scale (0 – Healthy, 1 – Mild-Moderate, 2 – Severe; criteria in Table 1) and, for a subset of 358 samples, the organism associated with the disease was identified by molecular diagnostics (Table 2), which detected *Pythium myriotylum* (n = 110), *Bacillus pumilus* (n = 98), *Fusarium solani* (n = 77) and *Fusarium oxysporum* (n = 73). The full dataset comprised 523 Healthy, 491 Mild-Moderate and 127 Severe observations (45.8%, 43.0% and 11.1% of samples respectively), reflecting a moderate class imbalance (4.1:1 between the largest and smallest class) typical of field disease surveys in which severely affected plants are comparatively rare.

Disease severity scoring

No international standard exists for scoring ginger rhizome rot, so severity was scored with a percentage-incidence and symptom-based scale of the kind commonly used for soft rot and wilt in rhizomatous crops. At each visit, the proportion of plants showing symptoms within the sampled plot was estimated and the condition of the rhizomes was inspected (Table 1). Healthy plants showed no foliar or pseudostem symptoms and had firm, turgid rhizomes of normal colour. The original Mild and Moderate grades were merged into one Mild-Moderate class, because most misclassifications in preliminary models occurred between these two adjacent grades. Severe plots showed wilting and collapse of pseudostems together with water-soaked, foul-smelling rhizome rot.

Table 1. Field criteria used to assign ginger rhizome disease severity classes

Class	Code	Field criterion
Healthy	0	No foliar or pseudostem symptoms; rhizomes firm, turgid and of normal colour; up to 5% of plants in the plot affected.
Mild-Moderate	1	Yellowing or curling of lower leaves, mild stunting, partial softening or discolouration of rhizomes; 6–50% of plants affected.
Severe	2	Wilting and collapse of pseudostems, water-soaked soft rot of rhizomes with foul odour; more than 50% of plants affected.

Soil sampling and laboratory analysis

At each visit, five cores were taken from the root zone (0–15 cm) of every plot and bulked. A fresh portion was used for moisture and temperature, and an air-dried portion, sieved through a 2 mm mesh, was used for chemical and textural analysis. The methods are listed in Table 2. Meters were calibrated before each sampling day with fresh buffer solutions (pH 4.01, 7.00 and 9.21) and a 0.01 M KCl standard (1.413 dS m^{-1} at 25 °C),

and sensor-based moisture readings were checked against gravimetric moisture determined on a subset of samples. The temperature probe was checked against a reference thermometer in an ice bath.

For molecular identification, DNA was extracted from symptomatic rhizome tissue, amplified with the primers given in Table 2, and the products were sequenced. Sequences were matched against GenBank by BLAST, and species-level identity was accepted at 98% or higher. Each PCR run included a no-template negative control and a positive control from a known isolate.

Table 2. Analytical methods used for the soil variables and for pathogen identification

Variable	Method	Reference
pH	1:2.5 soil:water suspension, glass-electrode pH meter	Jackson (1973)
EC	1:2.5 soil:water extract, conductivity meter	Jackson (1973)
Moisture	Gravimetric: oven-drying at 105 °C to constant mass	–
Temperature	Soil probe thermometer at 10 cm depth	–
Organic matter	Walkley–Black wet oxidation (organic carbon × 1.724)	Walkley & Black (1934)
Available N	Alkaline potassium permanganate method	Subbiah & Asija (1956)
Available P	Olsen extraction (0.5 M $NaHCO_3$, pH 8.5); molybdenum-blue colour read at 882 nm	Olsen et al. (1954)
Exchangeable K	1 M neutral ammonium acetate extraction; flame photometry	Hanway & Heidel (1952)
Sand, silt, clay	Hydrometer method; USDA texture triangle for texture class	Bouyoucos (1962)
Pathogen identification	DNA extraction (CTAB); PCR of the ITS region (ITS1/ITS4) and of the translation elongation factor 1-α gene (EF1/EF2) for *Fusarium*; 16S rRNA gene primers (27F/1492R) for bacteria; sequencing and BLAST matching against GenBank (≥ 98% identity)	White et al. (1990)

Feature engineering

Raw soil variables were supplemented with engineered features intended to capture non-linear and combined soil effects relevant to waterlogging-associated disease. Nitrogen, phosphorus and potassium concentrations, which were right-skewed, were log-transformed to stabilise variance. Interaction terms were constructed between moisture and EC and between moisture and clay content, on the premise that the disease-relevant effect of moisture is conditional on how readily the soil retains or dissipates water and dissolved salts. Composite and interaction features of this kind have proved informative in other soil-machine-learning applications: explainable and SHAP-based approaches have been used to reveal non-obvious combined effects of soil amendments on plant growth outcomes (Zhai et al., 2026) and to attribute nutrient-leaching risk to specific combinations of soil and fertiliser variables (He et al., 2025b). A composite drainage index was derived from the texture fractions and EC, and a nitrogen-to-potassium ratio was included as an indicator of nutrient balance. Because a majority of plots were sampled more than once, plot-level temporal features were also derived, namely the cumulative mean moisture recorded at a plot up to and including the current visit, the change in moisture and EC since the previous visit at the same plot, and an

indicator of whether a prior sample existed for that plot. Growth stage and USDA texture class were retained as categorical variables and one-hot encoded prior to model fitting. Pathogen identity and molecular confirmation status were excluded from the predictor set, as these are established only after, rather than before, disease expression and would not be available at the point an early-warning prediction is required. In total, 19 numeric and two categorical predictors (25 columns after one-hot encoding) entered the models.

Model development

Five supervised classifiers were trained to predict the three-class disease severity label: a Support Vector Machine (SVM) with a radial basis function kernel; a Random Forest; an Extreme Gradient Boosting model (XGBoost); a Light Gradient Boosting Machine (LightGBM); and a soft Voting Ensemble combining the probability outputs of the four base learners. Hybrid and stacked ensemble architectures combining multiple base learners have outperformed single algorithms in other soil- and yield-prediction tasks (Karbasi et al., 2025; Kaur et al., 2026), motivating the inclusion of a Voting Ensemble alongside individually trained classifiers in the present comparison, while multi-sensor feature fusion has likewise improved soil classification accuracy in related precision-agriculture applications (He et al., 2025a). All numeric predictors were standardised to zero mean and unit variance and categorical predictors were one-hot encoded, within a single preprocessing pipeline that was fitted on the training data only. Hyperparameters for the tree-based and kernel models were selected by randomised search (15 candidates for SVM and Random Forest, 10 for XGBoost and LightGBM) with three-fold group-wise cross-validation within the training partition (search ranges and selected values in Table A1), optimising macro-averaged F1-score so as not to favour the majority Healthy and Mild-Moderate classes at the expense of the minority Severe class, in keeping with integrated soil-health modelling frameworks that similarly prioritise balanced, degradation-sensitive performance metrics over raw accuracy (Khosravi Aqdam et al., 2026). The dataset was split by plot identity (GroupShuffleSplit, random seed 42) into a training partition (n = 910) and a held-out test partition (n = 231, approximately 20% of samples; 118 Healthy, 97 Mild-Moderate and 16 Severe), so that no plot appeared in both partitions. In the baseline configuration, SVM, Random Forest and LightGBM used balanced class weights and XGBoost used balanced sample weights during tuning; the Voting Ensemble combined the four tuned learners and was refitted without additional weighting.

Group-wise cross-validation

Because up to three soil samples were collected from the same plot at different growth stages, samples from a single plot are not independent: they share unmeasured plot-level characteristics, such as underlying drainage or historic management, that a model could exploit if the same plot appeared in both the training and validation partitions of a conventional random k-fold split. To avoid this source of optimistic bias, model performance was additionally assessed using three-fold GroupKFold cross-validation of the training partition, in which folds were constructed so that all samples belonging to a given plot were assigned to a single fold. This ensures that no plot contributes data to both the training and validation sets within a fold, providing a more conservative, and arguably more realistic, estimate of how each model would generalise to soil sampled from plots not previously seen. For the imbalance benchmark described below, three

repetitions of five-fold StratifiedGroupKFold cross-validation, again grouped by plot, were run on all 1,141 samples, so that all 127 Severe samples were scored out-of-fold.

Handling of class imbalance

Six strategies were compared for each classifier, keeping the tuned hyperparameters fixed: (i) no weighting; (ii) balanced class weights (balanced sample weights for XGBoost); (iii) SMOTE (Chawla et al., 2002); (iv) ADASYN (He et al., 2008), restricted to the Severe class; (v) Borderline-SMOTE (Han et al., 2005); and, for XGBoost only, (vi) a focal-loss objective (Lin et al., 2017). Oversampling was implemented with the imbalanced-learn package (Lemaître et al., 2017) inside the cross-validation pipeline, so that synthetic samples were generated from training folds only. The focal objective was a focal-weighted softmax loss ($\gamma = 2$, class-balanced weights) in which the modulating factor was treated as constant when computing the gradient, which is an approximation of the exact focal gradient. For the Voting Ensemble, the strategy was applied to every member. Strategies were compared on pooled out-of-fold predictions, and the uncertainty of differences in macro F1-score was estimated by bootstrap resampling of plots (2,000 resamples). Because the SVM was fitted with Platt-scaled probabilities, its class labels in this benchmark were taken as the arg-max of those probabilities.

Model interpretation

SHAP values (Lundberg & Lee, 2017) were computed with the tree explainer for the tuned XGBoost model on the test partition ($n = 231$) and are expressed in log-odds units. Dependence plots show how the value of a predictor translates into its contribution to the probability of Severe disease. To visualise the joint effect of moisture and EC, the mean predicted probability of Severe disease was calculated over a 25 × 25 grid spanning the 2nd to 98th percentiles of both variables, with the other predictors left at their observed values and the derived interaction terms recalculated. Local explanations were obtained with LIME (Ribeiro et al., 2016) for the first two Severe and the first Healthy samples of the test partition.

Spatial and temporal validation

Because the data come from one district and one season, two stricter internal checks were added. For spatial validation, plot centroids were grouped into five blocks by k-means clustering of their GPS coordinates, and each block was held out in turn. For temporal validation, each growth stage was held out in turn, and a forward test trained on early-vegetative samples and tested on mid-season and pre-harvest samples. Each model used its best imbalance strategy from the benchmark. The code also accepts an external data file, which will be used when data from another region or year become available.

Minimal sensor sets

Five nested feature sets were evaluated with Random Forest, XGBoost and LightGBM under repeated group-wise cross-validation: S1, moisture, log EC and their product; S2, S1 plus plot-level moisture history, EC and moisture changes and growth stage; S3, S2 plus sand, silt, clay, texture class, the Moisture×Clay term and the drainage index; S4, S3 plus pH, temperature and organic matter; and S5, all features. Model size was taken as the size of the serialised pipeline, and latency as the median time of 30 single-sample predictions on a CPU.

Statistical analysis of soil variables

Soil variables were compared between severity classes with the Kruskal–Wallis test, because several variables were skewed. The effect size was expressed as ε^2, and pairwise differences were tested with Mann–Whitney tests and Bonferroni correction. Associations among variables were summarised with Spearman rank correlation coefficients.

Model evaluation

Model performance on the held-out test set was summarised using overall accuracy, macro-averaged precision, recall and F1-score, weighted F1-score, quadratic weighted kappa (appropriate given the ordinal nature of the severity scale), and macro-averaged one-versus-rest receiver operating characteristic area under the curve (ROC-AUC). Balanced accuracy is reported where stated, and 95% intervals were obtained by bootstrap resampling of plots. Per-class precision, recall and F1-score were also computed to characterise performance for the minority Severe class specifically. Macro F1-score, defined as the unweighted mean of the per-class F1-scores, was adopted as the principal metric for cross-validation because it does not allow strong performance on the majority classes to mask weak performance on the minority class:

$$\text{Macro F1} = \frac{\text{F1}_{\text{Healthy}} + \text{F1}_{\text{Mild-Moderate}} + \text{F1}_{\text{Severe}}}{3} \quad (1)$$

where $\text{F1}_{\text{Healthy}}$, $\text{F1}_{\text{Mild-Moderate}}$ and $\text{F1}_{\text{Severe}}$ – F1-scores of the Healthy, Mild-Moderate and Severe classes, respectively.

Confusion matrices were examined to characterise the pattern of misclassification between adjacent and non-adjacent severity classes, and impurity- or gain-based feature importance (permutation importance for the SVM, which has no native impurity-based measure) was extracted from each trained model to identify the soil parameters most strongly associated with its predictions. SHAP and LIME were used for interpretation. All analyses were implemented in Python using standard open-source libraries (scikit-learn, Pedregosa et al., 2011; XGBoost, Chen & Guestrin, 2016; LightGBM, Ke et al., 2017), the hardware, software versions and settings are listed in Tables A1 and A2.

RESULTS AND DISCUSSION

Soil characteristics across disease severity classes

Table 3 shows how the 1,141 samples were distributed over growth stages and texture classes. The share of Severe samples rose slightly from the early vegetative stage (10.3%) to pre-harvest (12.0%), but varied more with texture: 18.6% of the clay soils were Severe, against 12.6% of the clay loams and 7.6% of the loamy soils. Healthy samples showed the opposite gradient (29.3%, 39.7% and 56.6%, respectively).

Table 4 summarises the soil physicochemical dataset by severity class. Kruskal–Wallis tests are reported because nitrogen, phosphorus, potassium and EC were strongly skewed; medians and interquartile ranges are therefore given instead of means. Moisture content and EC showed by far the strongest separation between classes ($\varepsilon^2 = 0.385$ and 0.251, respectively; both $P < 0.001$), rising monotonically from Healthy (median 28.2% moisture, 0.35 dS m^{-1}) through Mild-Moderate (35.8%, 0.44 dS m^{-1}) to Severe (40.8%, 0.60 dS m^{-1}).

Table 3. Number of samples in each disease severity class by growth stage and soil texture class (n = 1,141; 901 plots)

Group	Healthy	Mild-Moderate	Severe	Total
Growth stage				
Early vegetative	184	166	40	390
Mid-season (rhizome bulking)	167	166	42	375
Pre-harvest	172	159	45	376
Texture class				
Clay	41	73	26	140
Clay loam	198	238	63	499
Loamy	284	180	38	502
All samples	523	491	127	1,141

Table 4. Median [interquartile range] of soil physicochemical variables by disease severity class, with Kruskal–Wallis test statistics (n = 1,141)

Parameter	Healthy (n = 523)	Mild-Moderate (n = 491)	Severe (n = 127)	H	P	ε^2
pH	7.13 [6.56–7.66]	7.27 [6.71–7.70]	7.07 [6.63–7.71]	5.5	0.063	0.003
Moisture (%)	28.20 [24.40–32.05]	35.80 [31.70–39.85]	40.80 [36.40–44.15]	439.7	< 0.001	0.385
Temperature (°C)	27.80 [25.80–29.95]	29.10 [27.10–31.00]	28.70 [26.75–30.80]	34.7	< 0.001	0.029
Nitrogen N (mg kg^{-1})	12.60 [6.30–164.80]	13.40 [7.00–162.35]	30.40 [8.90–161.75]	7.4	0.025	0.005
Phosphorus P (mg kg^{-1})	12.00 [7.00–18.85]	12.80 [7.50–18.30]	15.20 [9.45–22.20]	14.0	< 0.001	0.011
Potassium K (mg kg^{-1})	260.00 [176.50–321.60]	284.00 [215.50–343.80]	325.10 [247.75–425.75]	45.8	< 0.001	0.038
Organic Matter (%)	1.79 [1.48–2.15]	1.85 [1.46–2.31]	1.96 [1.58–2.50]	10.8	0.004	0.008
EC (dS m^{-1})	0.35 [0.27–0.42]	0.44 [0.36–0.54]	0.60 [0.47–0.82]	287.6	< 0.001	0.251
Sand (%)	29.20 [22.10–38.00]	26.00 [21.10–32.00]	24.00 [19.60–29.80]	40.8	< 0.001	0.034
Silt (%)	33.20 [28.00–38.55]	32.00 [26.00–38.00]	31.20 [26.45–38.00]	6.4	0.041	0.004
Clay (%)	36.00 [28.00–44.00]	41.40 [32.00–48.00]	43.60 [36.00–49.70]	54.5	< 0.001	0.046

H: Kruskal–Wallis statistic (2 d.f.); ε^2: effect size. P-values are unadjusted; with Bonferroni correction across the 11 variables (α = 0.0045), the differences in pH, nitrogen and silt are not significant.

All three pairwise contrasts were significant for both variables. Clay, sand, potassium and temperature also differed between classes, but with small effect sizes ($\varepsilon^2 \leq 0.046$), whereas pH and silt content showed no reliable difference (P = 0.063 and 0.041 before correction for multiple testing, and no significant pairwise contrast for silt). Differences in nitrogen, phosphorus and organic matter were small ($\varepsilon^2 \leq 0.011$). Clay content alone correlated only weakly with EC across the full dataset (Spearman ρ = 0.02), indicating that clay's contribution to disease risk is better understood as modulating the effect of

moisture, through its influence on water retention, than as an independent driver; this is consistent with the prominence of the Moisture× Clay interaction term, rather than clay content alone, in the feature importance rankings reported below.

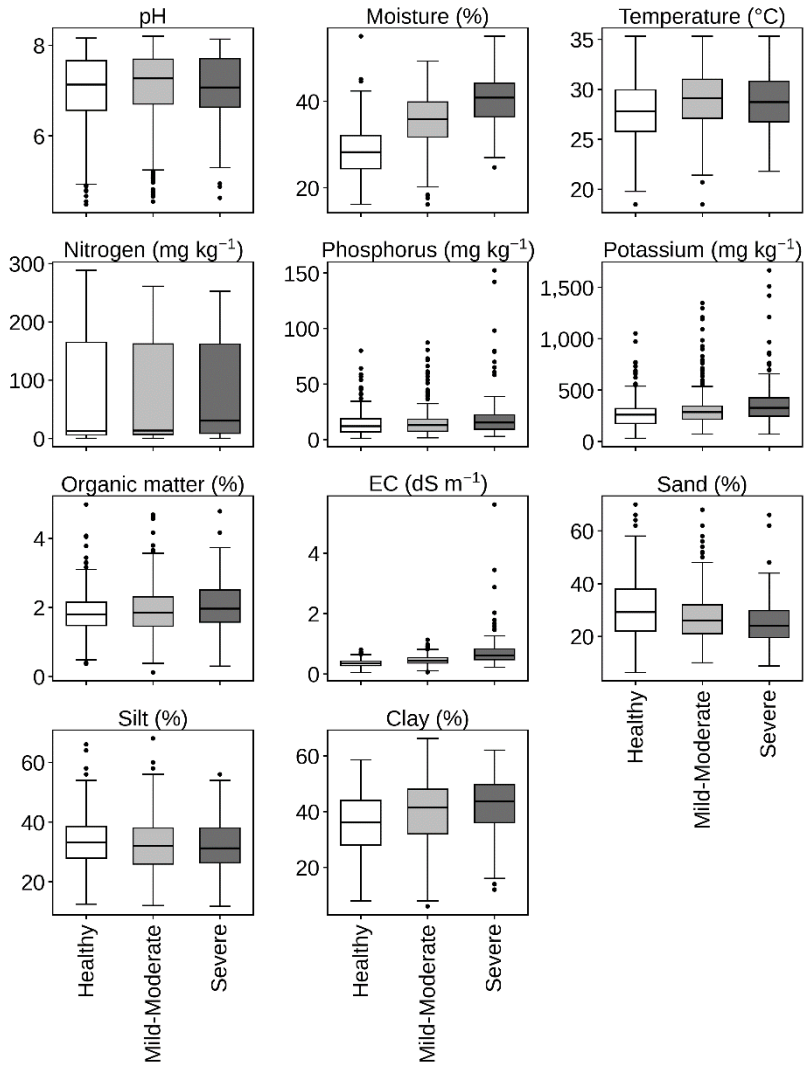

Figure 2. Distribution of the eleven soil physicochemical variables in the Healthy, Mild-Moderate and Severe classes (left to right within each panel). Boxes show the interquartile range with the median, whiskers extend to 1.5 times the interquartile range and dots are outliers (n = 1,141). Moisture and EC show the clearest increase with severity.

Figs 2 and 3 show the same pattern graphically. Moisture (Spearman ρ = 0.62 with severity) and EC (ρ = 0.50) were the variables most strongly associated with the severity score, followed by clay (0.22), potassium (0.19), sand (−0.19) and temperature (0.16). Moisture and EC were themselves moderately correlated (ρ = 0.42), as expected when more water mobilises more dissolved ions. The soils were close to neutral (median pH 7.1–7.3 in all classes), mostly clay loam or loamy in texture, and modest in organic

matter (median 1.8–2.0%). Potassium was positively related to clay ($\rho = 0.40$) and negatively to sand ($\rho = -0.40$), so its higher level in Severe plots probably reflects texture rather than a direct effect. The nitrogen distribution was very wide (interquartile range 6.3–164.8 mg kg^{-1} in the Healthy class) and showed little relation to severity ($\rho = 0.06$).

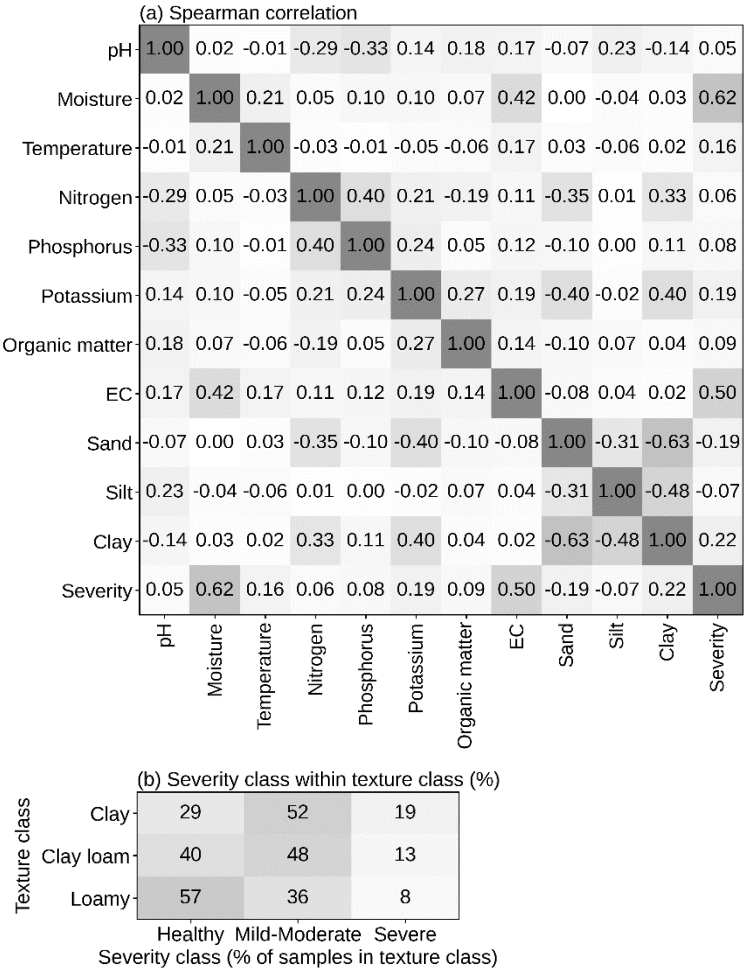

Figure 3. (a) Spearman rank correlation matrix of the soil variables and the ordinal severity score. (b) Percentage of Healthy, Mild-Moderate and Severe samples within each texture class; each row sums to 100%. Severe samples were proportionally most frequent in clay soils.

Samples in which a pathogen was identified had higher mean moisture (36.1–37.2%) and EC (0.48–0.55 dS m^{-1}) than samples in which none was identified (31.0% and 0.41 dS m^{-1}); the highest EC (0.55 dS m^{-1}), phosphorus (18.3 mg kg^{-1}) and potassium (353 mg kg^{-1}) were recorded where *Pythium myriotylum* was found. These associations describe the conditions under which disease was observed and do not by themselves prove cause and effect.

This moisture- and EC-dominated separation between disease classes is consistent with a broader pattern in the soil-machine-learning literature, in which moisture and its

electrochemical correlates are repeatedly identified as the most tractable and informative soil variables to model. Machine learning has been used to invert and analyse the sensitivity of soil moisture in other intensively irrigated horticultural crops (Wang et al., 2025b), to select the most relevant predictor subset for soil moisture prediction (Katipoğlu et al., 2025), and to track soil moisture reliably at continental scale from satellite data (Asadollah et al., 2026). Comparable UAV- and multispectral-based approaches have estimated fine-scale soil water content and linked it directly to crop outcomes in cotton (Li et al., 2025) and wheat (Yang et al., 2026), while convolutional architectures applied to ground-penetrating radar data have extended moisture estimation to multiple soil depths (Vahidi et al., 2025). Reviews integrating remote sensing with artificial intelligence more generally confirm that soil moisture remains the single most frequently and successfully modelled soil property across this literature (P et al., 2025). The salinity side of this relationship has received parallel attention, with machine learning used to map and risk-zone soil salinisation at basin scale (Wang et al., 2026) and, closer to the present study's use of pH as a co-predictor, to spatially map soil profile pH in saline-alkali farmland using stacking ensemble models with explicit spatial interpretability analysis (Lei et al., 2026). Taken together, this body of work supports treating moisture and EC, rather than the full soil-chemical panel, as the primary tractable signal in soil-based predictive modelling, which is precisely the pattern observed for ginger disease severity in the present dataset.

Comparative model performance

Table 5 reports held-out test-set performance for the five classifiers. Accuracy was almost identical across models (74.5–75.8%), and XGBoost achieved the highest macro F1-score (0.696), ahead of Random Forest (0.684), SVM (0.664), the Voting Ensemble (0.653) and LightGBM (0.644). XGBoost also had the highest macro recall (0.714).

Table 5. Held-out test-set performance of five classification algorithms for three-class ginger disease severity prediction (n = 231)

Model	Accuracy	Macro P	Macro R	Macro F1	Weighted F1	QWK	ROC-AUC
SVM	0.753	0.651	0.692	0.664	0.756	0.674	0.880
Random Forest	0.758	0.689	0.680	0.684	0.756	0.682	0.883
XGBoost	0.758	0.684	0.714	0.696	0.757	0.683	0.879
LightGBM	0.745	0.657	0.635	0.644	0.742	0.658	0.863
Voting Ensemble	0.758	0.666	0.643	0.653	0.754	0.676	0.884

Macro P, Macro R: macro-averaged precision and recall; QWK: quadratic weighted kappa; ROC-AUC: macro-averaged one-versus-rest area under the receiver operating characteristic curve.

The Voting Ensemble, Random Forest and SVM obtained the highest macro ROC-AUC values (0.884, 0.883 and 0.880), reflecting good class-probability ranking, but this was not accompanied by the best per-class balance, as discussed below. Quadratic weighted kappa was 0.66–0.68 for all models. Because the test set contains only 231 samples from a limited number of plots, the model differences are small relative to their uncertainty: plot-level bootstrap 95% intervals for macro F1-score were 0.584–0.778 for Random Forest, 0.610–0.779 for XGBoost and 0.549–0.739 for LightGBM, so the ranking should be read as indicative only.

Table 6 reports the corresponding three-fold GroupKFold cross-validated macro F1-scores obtained within the training partition. The differences between models were again small. Random Forest achieved the highest mean macro F1-score (0.664), followed by XGBoost (0.657), the Voting Ensemble (0.654), LightGBM(0.652) and SVM (0.650); LightGBM was the most stable (standard deviation 0.006) and SVM and XGBoost the least (0.017). This ordering differs from that on the test set, a useful reminder that performance rankings from one test partition can be sensitive to which plots happen to fall within it, particularly for the Severe class, which contributed only 16 of the 231 test-set samples. We consider the cross-validated estimates the more trustworthy indicator of likely generalisation to unseen plots, while noting that XGBoost's competitive performance under both schemes, together with its best per-class F1-score for the Severe class (see below), supports its selection as the primary candidate model, with Random Forest as a robust alternative where evaluation stability across plots is prioritised.

Table 6. Three-fold GroupKFold cross-validated macro F1-score (grouped by plot identity) for five classification algorithms

Model	CV Macro F1 (mean)	CV Macro F (SD)
SVM	0.650	0.017
Random Forest	0.664	0.014
XGBoost	0.657	0.017
LightGBM	0.652	0.006
Voting Ensemble	0.654	0.011

Class-wise diagnostic performance and confusion patterns

Table 7 presents precision, recall and F1-score for each severity class and model, and Fig. 4 shows the corresponding confusion matrices. Across all five models, most misclassification occurred between the adjacent Healthy and Mild-Moderate classes, while confusion between the extreme Healthy and Severe classes was rare (at most one sample in either direction per model), which is reassuring from an operational standpoint, since the costliest error for a disease early-warning system is failing to flag an actually diseased plot as healthy rather than misjudging its severity by one grade.

Table 7. Class-wise precision, recall and F1-score of five classification algorithms on the held-out test set (n = 231)

Model	Class	Precision	Recall	F1-score	Support
SVM	Healthy	0.829	0.864	0.846	118
	Mild-Moderate	0.750	0.649	0.696	97
	Severe	0.375	0.562	0.450	16
Random Forest	Healthy	0.818	0.839	0.828	118
	Mild-Moderate	0.716	0.701	0.708	97
	Severe	0.533	0.500	0.516	16
XGBoost	Healthy	0.813	0.847	0.830	118
	Mild-Moderate	0.739	0.670	0.703	97
	Severe	0.500	0.625	0.556	16
LightGBM	Healthy	0.811	0.839	0.825	118
	Mild-Moderate	0.698	0.691	0.694	97
	Severe	0.462	0.375	0.414	16
Voting Ensemble	Healthy	0.816	0.864	0.840	118
	Mild-Moderate	0.720	0.691	0.705	97
	Severe	0.462	0.375	0.414	16

Recall for the minority Severe class varied considerably between algorithms: XGBoost correctly identified 10 of 16 Severe cases (recall 0.625), the highest of any model, with precision 0.500 and the highest F1-score for the class (0.556); SVM identified 9 of 16 (recall 0.562) but with the lowest precision (0.375), reflecting false Severe alarms among Mild-Moderate samples. Random Forest identified 8 of 16 (0.500, F1-score 0.516). Notably, the Voting Ensemble, despite strong aggregate accuracy and the best macro ROC-AUC, and LightGBM both showed the weakest Severe-class recall of all five models (0.375, six of sixteen correctly identified), suggesting that averaging probabilities across base learners can suppress correct identification of a small minority class even when it improves ranking-based metrics overall. This trade-off has practical implications: an operational system prioritising sensitivity to severe disease, for example as a first-pass screening layer that flags candidate plots for closer inspection, may reasonably favour XGBoost or SVM over the Voting Ensemble, notwithstanding the latter's higher aggregate accuracy. Because the Severe class contains only 16 test samples, one reclassified plot changes its recall by 0.0625, and the bootstrap 95% intervals for Severe recall are wide (Random Forest 0.25–0.76, XGBoost 0.36–0.87, LightGBM 0.14–0.65). The pooled out-of-fold analysis in the next section, which scores all 127 Severe samples, was added to address this limitation.

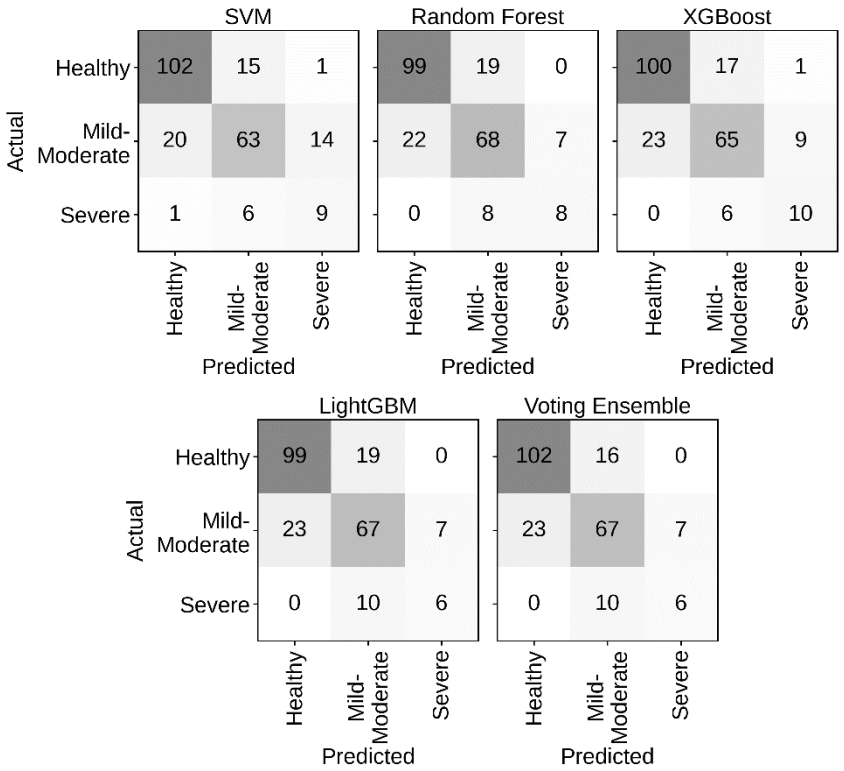

Figure 4. Confusion matrices for five classification algorithms predicting three-class ginger disease severity on the held-out test set ($n = 231$; 118 Healthy, 97 Mild-Moderate, 16 Severe). Rows are actual classes and columns are predicted classes; cell values are sample counts and shading is proportional to the count. Most errors fall between adjacent classes.

Effect of class-imbalance handling

Table 8 and Fig. 5 compare the six imbalance strategies on pooled out-of-fold predictions for all 1,141 samples (three repetitions, standard deviation of macro F1-score at most 0.015), which include all 127 Severe samples. Three findings stand out. First, no rebalancing strategy improved macro F1-score over the unweighted models. The best setting by macro F1-score was no weighting for SVM (0.686), XGBoost (0.678), LightGBM (0.676) and the Voting Ensemble (0.686), and balanced class weights for Random Forest (0.671). Plot-level bootstrap intervals for the difference between the best setting and its comparator included zero for Random Forest (+0.012; 95% CI −0.015 to 0.041), LightGBM (+0.013; −0.002 to 0.029) and the Voting Ensemble (+0.002; −0.016 to 0.017), while removing class weights improved macro F1-score modestly but reliably for SVM (+0.034; 0.010 to 0.060) and XGBoost (+0.027; 0.003 to 0.050).

Table 8. Effect of class-imbalance strategy on pooled out-of-fold performance (three repetitions of five-fold StratifiedGroupKFold, n = 1,141; 127 Severe samples). Values are means over repetitions

Model	Strategy	Accuracy	Macro F1	ROC-AUC	Severe precision	Severe recall	Severe F1
SVM	None	0.752	0.686	0.872	0.777	0.381	0.510
	Class weight	0.739	0.641	0.864	0.639	0.278	0.387
	SMOTE	0.713	0.653	0.853	0.456	0.499	0.476
	ADASYN	0.714	0.654	0.849	0.451	0.517	0.482
	Borderline-SMOTE	0.709	0.645	0.848	0.445	0.467	0.456
Random Forest	None	0.744	0.665	0.864	0.809	0.323	0.461
	Class weight	0.732	0.671	0.864	0.571	0.446	0.501
	SMOTE	0.711	0.656	0.858	0.458	0.541	0.496
	ADASYN	0.689	0.634	0.853	0.405	0.567	0.473
	Borderline-SMOTE	0.699	0.641	0.854	0.420	0.535	0.471
XGBoost	None	0.746	0.678	0.868	0.723	0.381	0.498
	Class weight	0.716	0.655	0.862	0.458	0.504	0.480
	SMOTE	0.709	0.655	0.858	0.448	0.556	0.496
	ADASYN	0.698	0.641	0.857	0.408	0.564	0.474
	Borderline-SMOTE	0.699	0.638	0.855	0.400	0.535	0.458
	Focal loss	0.723	0.662	0.862	0.509	0.472	0.490
LightGBM	None	0.742	0.676	0.850	0.701	0.391	0.502
	Class weight	0.736	0.671	0.850	0.613	0.415	0.494
	SMOTE	0.721	0.662	0.846	0.527	0.472	0.497
	ADASYN	0.717	0.656	0.848	0.492	0.483	0.487
	Borderline-SMOTE	0.718	0.656	0.846	0.492	0.470	0.480
Voting Ensemble	None	0.753	0.686	0.873	0.767	0.386	0.513
	Class weight	0.743	0.679	0.869	0.617	0.430	0.507
	SMOTE	0.725	0.665	0.864	0.502	0.488	0.495
	ADASYN	0.722	0.661	0.862	0.475	0.507	0.490
	Borderline-SMOTE	0.720	0.660	0.861	0.481	0.493	0.487

SVM class labels were taken as the arg-max of Platt-scaled probabilities in this benchmark; the SVM results should be interpreted with that decision rule in mind.

Second, rebalancing moved the operating point along the precision–recall trade-off for the Severe class rather than improving it overall. Recall for the Severe class rose from 0.32–0.39 without weighting to 0.47–0.57 with SMOTE, ADASYN or Borderline-

SMOTE (highest for Random Forest with ADASYN, 0.567, and XGBoost with ADASYN or SMOTE, 0.564 and 0.556), while precision fell from 0.70–0.81 to 0.40–0.53. The F1-score of the Severe class stayed between 0.46 and 0.51 in every combination, and overall accuracy fell by about 2 to 5 percentage points. The focal-loss XGBoost model (macro F1-score 0.662, Severe precision 0.509, recall 0.472) fell between the unweighted and the class-weighted XGBoost models and offered no gain.

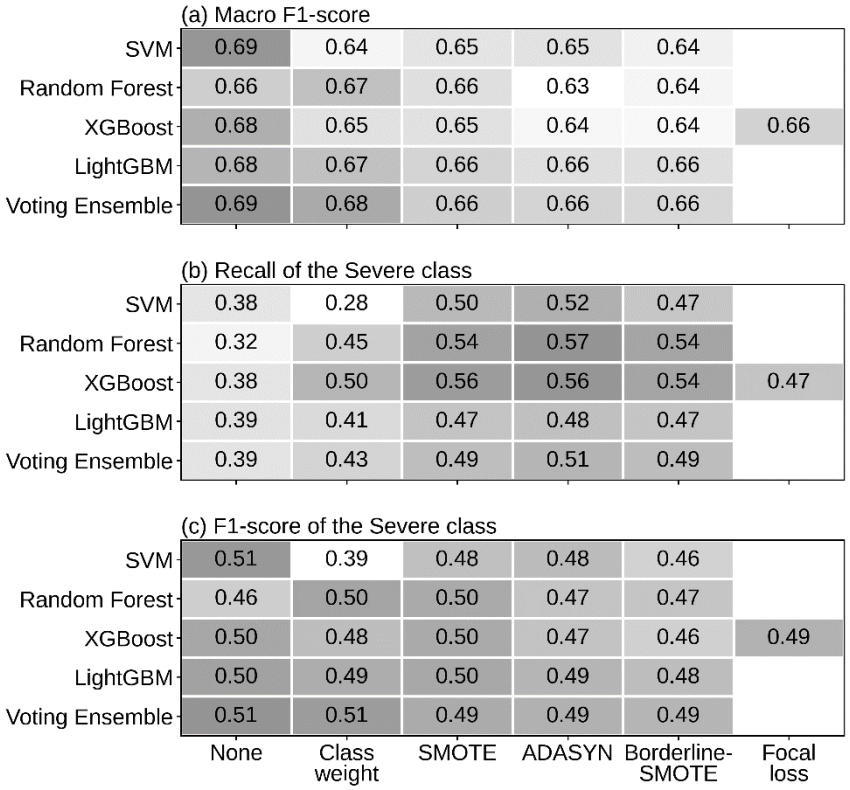

Figure 5. Heat maps of (a) macro F1-score, (b) recall of the Severe class and (c) F1-score of the Severe class for each classifier and imbalance strategy on pooled out-of-fold predictions (n = 1,141; 127 Severe samples). Empty cells indicate that the focal-loss objective was applied to XGBoost only. Oversampling increases Severe recall (b) without raising macro F1-score (a) or Severe F1-score (c).

Third, for the Voting Ensemble, which might be expected to benefit most from rebalancing, balancing members with class weights raised Severe recall from 0.386 to 0.430, and SMOTE, ADASYN or Borderline-SMOTE raised it to 0.488–0.507. However, macro F1-score changed from 0.686 to 0.679 with weights and to 0.660–0.665 with oversampling, because precision fell from 0.767 to 0.475–0.502. Imbalance handling therefore improved the sensitivity of the ensemble but not its balanced score. The limited gain is consistent with class overlap rather than a shortage of minority samples: in the pooled confusion matrices (Fig. 6), 55–61% of Severe samples were labelled Mild-Moderate, whereas only 1 of 127 was labelled Healthy and at most 4 of 523 Healthy samples were labelled Severe. For screening, in which a missed Severe plot

is costly, Random Forest or XGBoost trained with SMOTE or ADASYN (recall about 0.55–0.57, precision about 0.41–0.46) is therefore the better choice, whereas the unweighted ensemble produces fewer false alarms (precision 0.77) but misses about six in ten Severe plots.

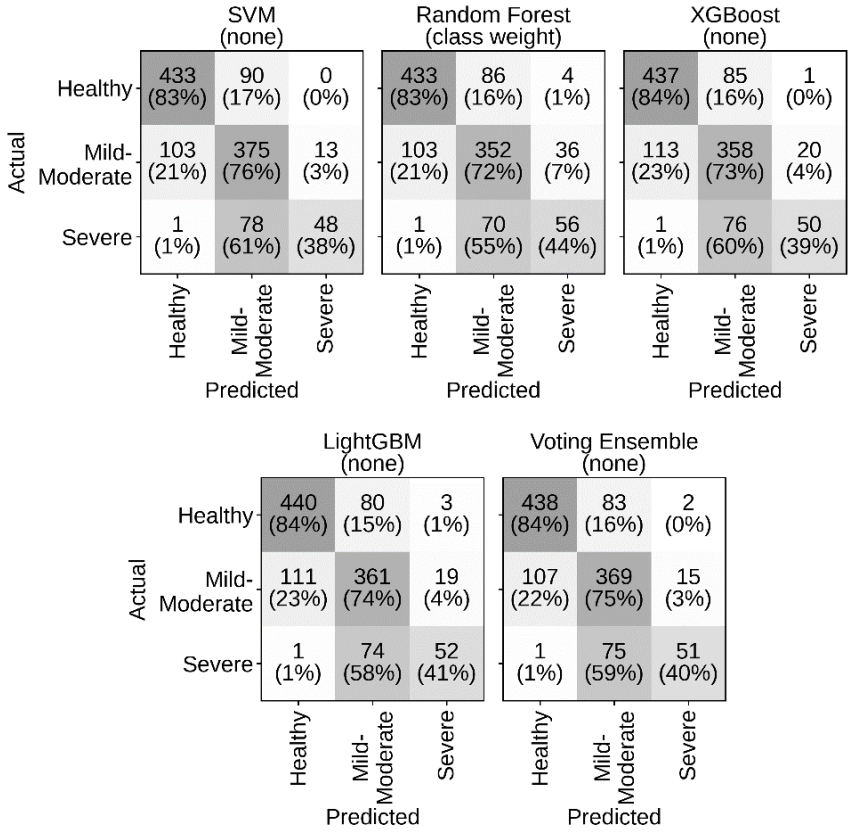

Figure 6. Confusion matrices from pooled out-of-fold predictions (first repetition; $n = 1{,}141$, including all 127 Severe samples) for each classifier with its best imbalance strategy (shown in brackets). Cell values are counts with row percentages in parentheses. Severe samples were rarely classed as Healthy.

Feature importance and agronomic interpretation

Fig. 7 and Table 9 summarise feature importance for the four models for which a native or model-agnostic importance measure was available. For Random Forest, XGBoost and SVM, moisture-related features dominated the rankings: the plot-level cumulative moisture mean, the Moisture×EC interaction term, raw moisture content and the Moisture×Clay interaction term occupied four of the top five positions for Random Forest, together with log-transformed EC, and the five features accounted for about 60% of the total Gini importance. In XGBoost, the cumulative moisture mean, Moisture×EC, log EC and Moisture×Clay were again among the top five, together with a loamy-texture indicator, which agrees with the lower share of Severe samples in loamy soils (Table 3). SVM permutation importance ranked log-transformed EC clearly first (0.111), followed

by the cumulative moisture mean, Moisture×EC and raw moisture, with the remaining variables contributing little. LightGBM's split-based importance measure produced a different leading order, ranking Moisture×Clay, organic matter, pH and temperature among the top five alongside Moisture×EC; however, raw moisture content ranked sixth and the moisture-related terms overall remained prominent, indicating that the divergence reflects differences in how split-count-based importance is computed relative to Gini- or permutation-based measures, rather than a genuine disagreement between algorithms about the underlying soil-disease relationship. Native importances do not show the direction of an effect; this is addressed with SHAP and LIME below.

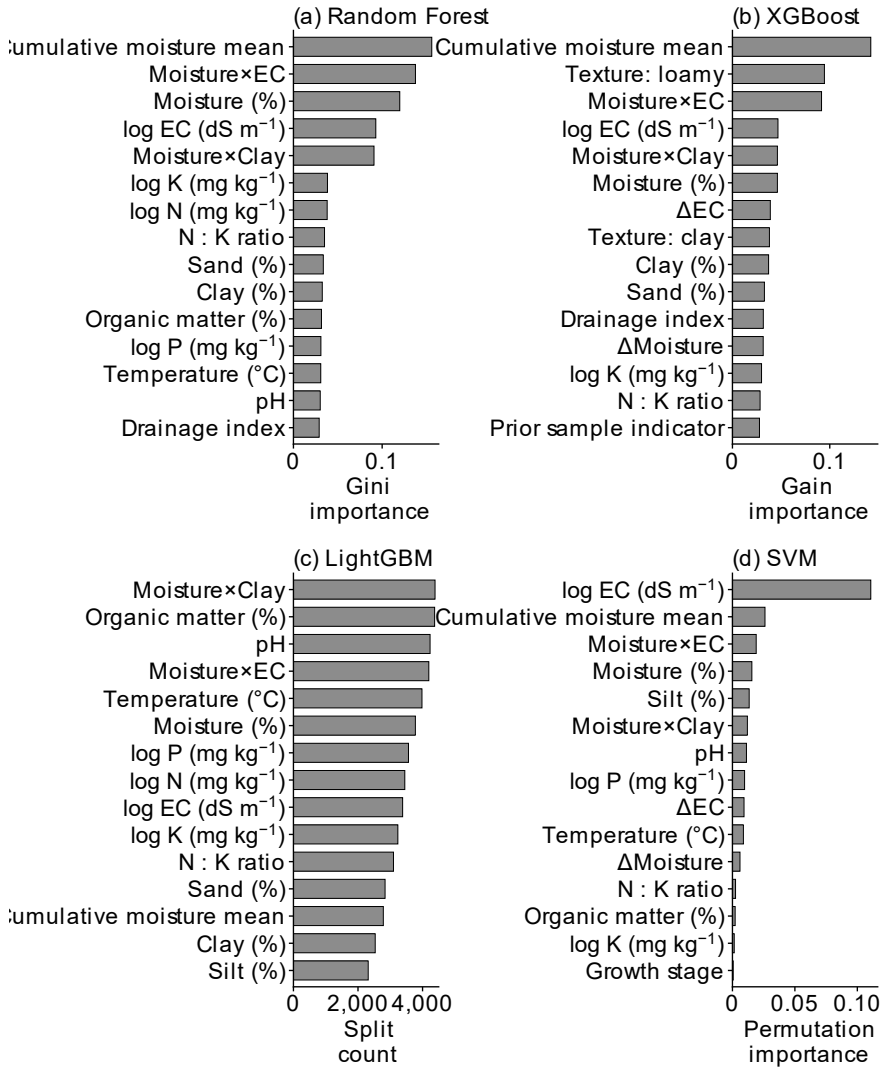

Figure 7. Relative feature importance of the fifteen most influential predictors of ginger disease severity for Random Forest, XGBoost, LightGBM and SVM (Gini, gain, split-count and permutation importance, respectively; the horizontal axes differ in scale and are not comparable between panels). Moisture-related predictors rank highest in three of the four models.

Table 9. Five most influential predictors of ginger disease severity for four classification algorithms, ranked by model-specific feature importance

Rank	Random Forest	XGBoost	LightGBM	SVM
1	Cumulative moisture mean (0.156)	Cumulative moisture mean (0.142)	Moisture×Clay (4,387)	log EC (0.111)
2	Moisture×EC (0.138)	Loamy texture (0.095)	Organic matter (4,374)	Cumulative moisture mean (0.026)
3	Moisture (0.120)	Moisture×EC (0.092)	pH (4,233)	Moisture×EC (0.019)
4	log EC (0.093)	log EC (0.047)	Moisture×EC (4,194)	Moisture (0.016)
5	Moisture×Clay (0.091)	Moisture×Clay (0.046)	Temperature (3,982)	Silt (0.013)

Values in parentheses are relative feature importance (Gini importance for Random Forest; gain-based importance for XGBoost; split-count importance for LightGBM; permutation importance for SVM); the underlying units differ between columns and are not directly comparable in magnitude across models.

The comparatively minor role of the nutrient variables (nitrogen, phosphorus, potassium, organic matter) relative to moisture and EC in the present results should not be read as evidence that nutrient status is unimportant to ginger soil management generally; rather, it reflects the specific outcome being modelled. Machine learning has proved highly capable of estimating soil organic matter (Cui et al., 2025) and total nitrogen content (Dong et al., 2025) directly from physicochemical and spectral data, and of characterising nitrogen and phosphorus buffering capacity at landscape scale (Cao et al., 2026), underlining that these nutrients remain highly predictable and agronomically important soil properties in their own right, and that comparative algorithm benchmarking of the kind reported here, including comparison against semi-mechanistic and conventional fertiliser-recommendation approaches (Asamoah et al., 2025), is equally applicable to a nutrient-status prediction task; it is only for the specific outcome of rhizome disease severity that moisture and EC emerge as the dominant, rather than merely contributing, signal.

The consistent prominence of moisture, EC and their interaction across algorithms is agronomically coherent. Zoospores of *Pythium* species, the most frequently molecularly confirmed pathogen in this dataset, are motile only in free water and require saturated or near-saturated soil conditions to reach and infect rhizome tissue; elevated EC under these conditions is a plausible correlate of restricted drainage and localised salt or nutrient accumulation rather than a direct causal agent in its own right. The comparatively minor independent contribution of clay content, despite clay's role in water retention, together with the strong showing of the Moisture×Clay interaction term, suggests that soil texture acts on disease risk primarily by modulating moisture behaviour rather than through any separate mechanism.

Direction of effects: SHAP and LIME

SHAP analysis of the XGBoost model on the test set confirmed the importance ranking and showed how predictor values change the predicted probability of Severe disease (Table 10; Figs 8–9). The cumulative moisture mean (mean absolute SHAP value 0.430), Moisture×EC (0.375) and Moisture×Clay (0.200) were the three most influential features overall, followed by raw moisture (0.148) and log EC (0.138).

Table 10. Mean absolute SHAP value (log-odds) of the eight most influential features of the XGBoost model on the test set (n = 231), overall and by class

Feature	Overall	Healthy	Mild-Moderate	Severe
Cumulative moisture mean	0.430	0.689	0.073	0.528
Moisture×EC	0.375	0.423	0.175	0.527
Moisture×Clay	0.200	0.197	0.069	0.333
Moisture (%)	0.148	0.105	0.195	0.145
log EC	0.138	0.188	0.033	0.192
Clay (%)	0.087	0.089	0.034	0.138
Sand (%)	0.083	0.130	0.021	0.098
Organic matter (%)	0.080	0.103	0.098	0.038

For the Severe class, the same three terms dominated (0.528, 0.527 and 0.333), whereas nitrogen (0.051) and phosphorus (0.035) contributed little and potassium contributed moderately (0.110), probably as a proxy for clay content.

The beeswarm plot (Fig. 8) shows that high values of the cumulative moisture mean, Moisture×EC, Moisture×Clay, log EC, raw moisture, clay and potassium push the prediction towards Severe disease, whereas high sand content pushes it away.

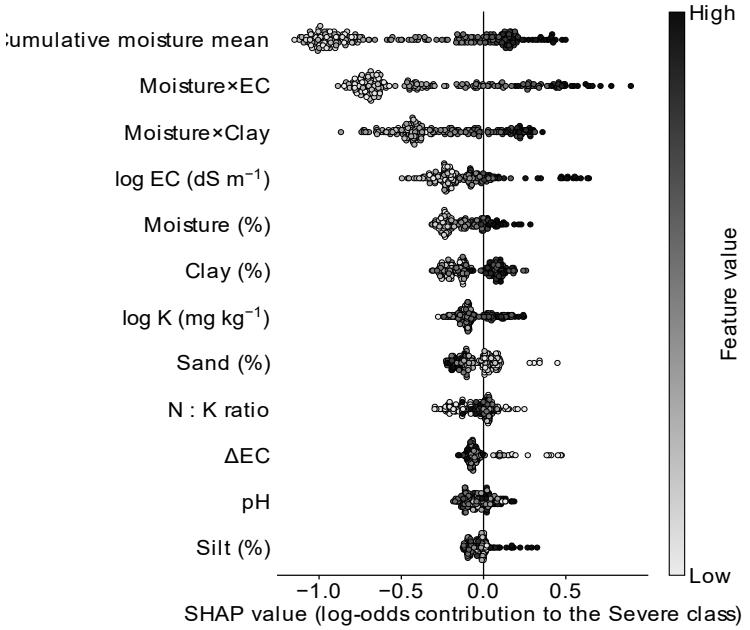

Figure 8. SHAP beeswarm plot for the Severe class of the XGBoost model (test set, n = 231). Each point is one sample; the horizontal position is the SHAP value (log-odds contribution to the Severe class) and the shading shows the standardised feature value (dark, high; light, low).

The dependence plots (Fig. 9) show that the contribution of each predictor is not linear: the SHAP value of Moisture×EC rises from about −0.75 at low values to between +0.5 and +0.9 at the highest standardised values, and the contribution of log EC jumps to about +0.5 for the highest EC values. At the same moisture, samples with higher EC lie slightly above those with lower EC, which indicates that EC adds risk on top of wet conditions.

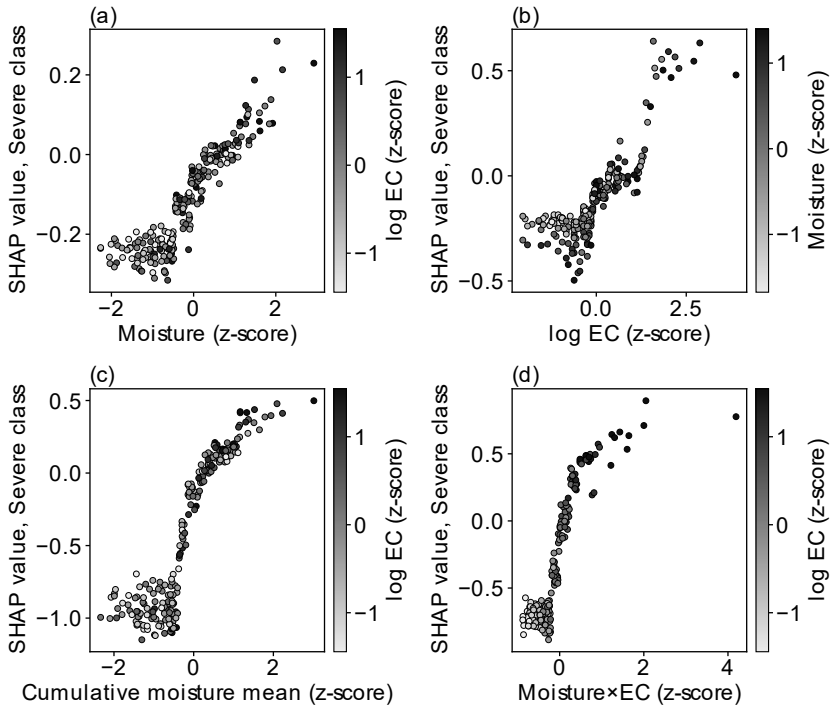

Figure 9. SHAP dependence plots for the Severe class: contribution of (a) moisture, (b) log EC, (c) cumulative moisture mean and (d) Moisture×EC. Feature values on the horizontal axes are standardised (z-scores). The shading shows the interacting variable (log EC in a, c and d; moisture in b; dark, high).

The response surface in Fig. 10 puts these effects in field units: the mean predicted probability of Severe disease remains low (below about 0.12) where moisture is below roughly 30% and EC below about 0.5 dS m^{-1}, increases with either variable, and exceeds 0.5 only in the upper-right corner, where moisture is above about 40% and EC approaches 0.8–1.0 dS m^{-1}. In other words, high moisture combined with high EC is the condition under which the model assigns the greatest probability to Severe rhizome disease.

LIME explanations for individual test samples agreed with the global picture (Fig. 11).

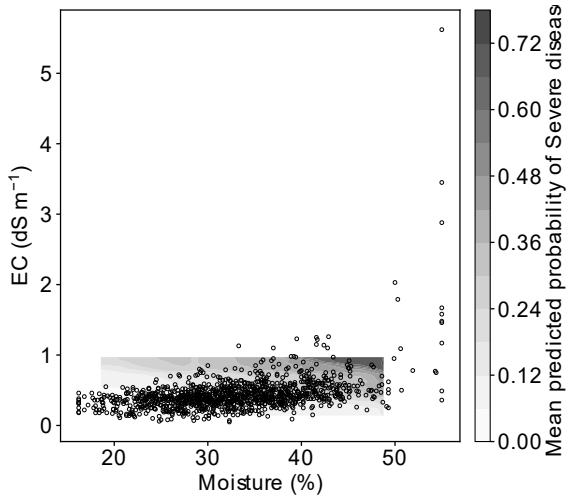

Figure 10. Mean predicted probability of Severe disease from the XGBoost model over a grid of soil moisture and EC. The coloured surface covers the 2nd to 98th percentile range of both variables; points are the observed samples.

For both Severe samples, the three largest contributions towards the Severe class came from a high Moisture×EC value (about 0.21), a high cumulative moisture mean (about 0.13) and a high EC (0.08–0.09), with a small contribution from raw moisture. The few opposing contributions were small (mid-range clay and potassium in the first sample; high sand and low nitrogen in the second). For the Healthy sample, a mid-range cumulative moisture mean and a mid-range Moisture×EC value were the main features supporting the Healthy class, and their contributions were smaller (about 0.07 each). Because LIME fits a local linear surrogate, these explanations describe single predictions and should not be generalised.

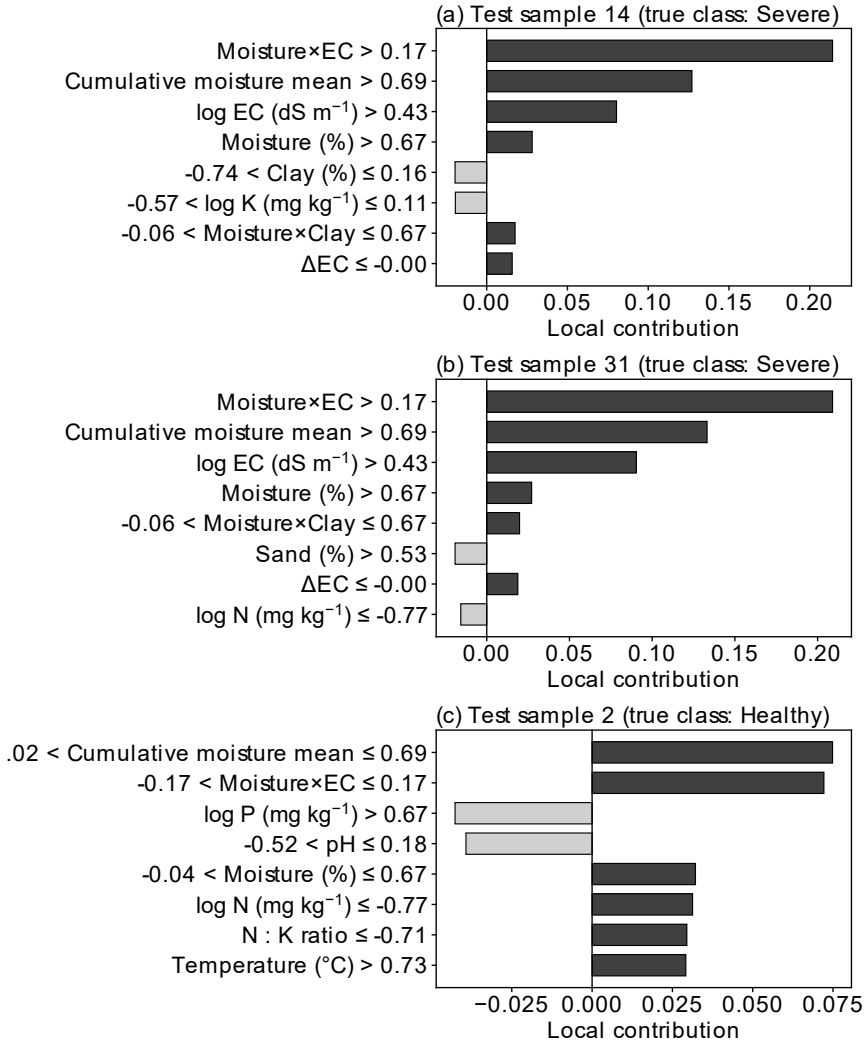

Figure 11. LIME explanations for three test samples: (a, b) two Severe samples (explaining the Severe class) and (c) one Healthy sample (explaining the Healthy class). Bars show the local contribution of each feature value range (standardised units); dark bars support and light bars oppose the explained class.

Spatial and temporal generalisation

The internal validation checks are summarised in Table 11 and Fig. 12. Under leave-one-block-out spatial validation (five blocks of 157–316 samples, 21–33 Severe samples each), macro F1-scores were 0.679–0.696 and accuracy 0.733–0.763 for all models, which is similar to the pooled cross-validation results (0.67–0.69 for the best settings).

Table 11. Spatial (leave-one-GPS-block-out, 5 folds) and temporal (leave-one-growth-stage-out plus forward test, 4 folds) validation. Values are the mean ± standard deviation across folds

Model	Accuracy	Macro F1	QWK	Severe recall
Temporal (growth stage)				
SVM	0.736 ± 0.038	0.644 ± 0.037	0.641 ± 0.063	0.309 ± 0.068
Random Forest	0.725 ± 0.032	0.671 ± 0.025	0.673 ± 0.042	0.491 ± 0.056
XGBoost	0.730 ± 0.024	0.654 ± 0.029	0.645 ± 0.039	0.343 ± 0.075
LightGBM	0.720 ± 0.037	0.645 ± 0.056	0.635 ± 0.059	0.358 ± 0.117
Voting Ensemble	0.740 ± 0.025	0.664 ± 0.032	0.661 ± 0.038	0.358 ± 0.100
Spatial (GPS block)				
SVM	0.752 ± 0.018	0.690 ± 0.022	0.691 ± 0.018	0.394 ± 0.049
Random Forest	0.733 ± 0.028	0.682 ± 0.025	0.679 ± 0.029	0.491 ± 0.040
XGBoost	0.749 ± 0.030	0.679 ± 0.042	0.680 ± 0.041	0.379 ± 0.076
LightGBM	0.749 ± 0.026	0.682 ± 0.031	0.681 ± 0.037	0.397 ± 0.046
Voting Ensemble	0.763 ± 0.025	0.696 ± 0.030	0.696 ± 0.041	0.388 ± 0.034

Under temporal validation (test sets of 375–751 samples with 40–87 Severe samples), macro F1-scores were somewhat lower and more variable (0.644–0.671), with Random Forest the most stable (0.671 ± 0.025). Models therefore did not collapse when they were applied to plots in unseen areas or to growth stages not used for training. Recall for the Severe class was lower than the overall pooled values (0.31–0.49 in the temporal and 0.38–0.49 in the spatial tests), with Random Forest the best in both schemes (0.491). These checks are internal, because all data come from one district and one season; they reduce, but do not remove, the need for validation on an external region or year, which we identify as the main limitation of the present work.

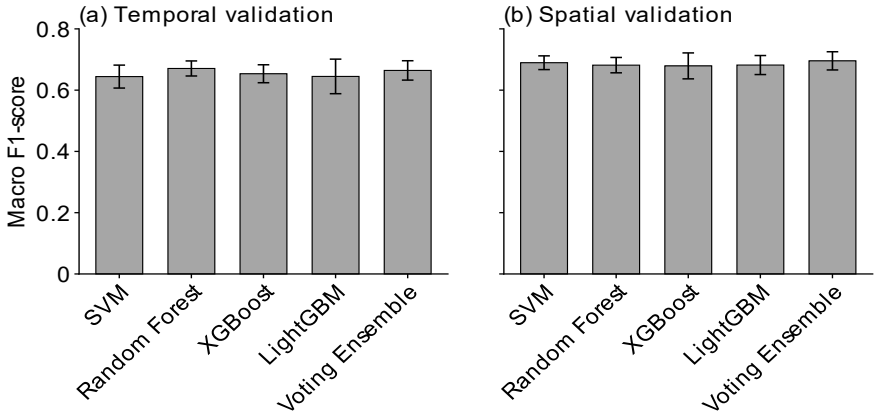

Figure 12. Macro F1-score of the five classifiers under temporal (leave-one-growth-stage-out and forward) and spatial (leave-one-GPS-block-out) validation. Bars are means across folds and error bars are standard deviations.

Implications for low-cost, field-deployable monitoring

The explicit purpose of this work was to identify what a field-deployable monitoring tool needs to measure. Table 12 and Fig. 13 show how model performance changed as sensors were added. A set built only from moisture and EC (S1: moisture, log EC and their product) reached a macro F1-score of 0.566–0.602, which is 84–90% of the score of the full feature set (0.655–0.671), and its Severe-class recall was as high as with all features for XGBoost (0.551 against 0.504). Adding moisture history and growth stage (S2) brought no gain. The largest further step came from soil texture (S3, +0.04 to +0.06), followed by pH, temperature and organic matter (S4, +0.02 to +0.03), whereas adding nitrogen, phosphorus and potassium (S5) improved the macro F1-score by at most 0.02 and lowered it for XGBoost. A moisture-and-EC sensor pair, optionally supplemented by a one-off soil texture determination, is therefore the most cost-effective configuration. The three-feature XGBoost model was 842 KB and needed a median of 2.6 ms per prediction on the notebook CPU, compared with 3,480 KB and 14.5 ms for Random Forest and 6,045 KB and 3.2 ms for LightGBM. These values were measured on a standard CPU, not on a microcontroller, and models of this size would need conversion or compression before running on a TinyML device.

Table 12. Performance and size of models built on nested sensor sets (repeated five-fold group-wise cross-validation, n = 1,141). S1: moisture, log EC, Moisture×EC; S2: S1 + moisture and EC history, growth stage; S3: S2 + texture variables; S4: S3 + pH, temperature, organic matter; S5: all features

Set	*n*	Model	Macro F1	Severe recall	Accuracy	Size (KB)	Latency (ms)
S1	3	Random Forest	0.602 ± 0.007	0.472	0.657	3,480	14.5
		XGBoost	0.592 ± 0.006	0.551	0.645	842	2.6
		LightGBM	0.566 ± 0.008	0.394	0.624	6,045	3.2
S2	8	Random Forest	0.606 ± 0.005	0.449	0.669	3,281	25.1
		XGBoost	0.592 ± 0.008	0.517	0.649	833	3.9
		LightGBM	0.564 ± 0.003	0.344	0.632	6,066	4.6
S3	14	Random Forest	0.648 ± 0.006	0.478	0.704	3,279	16.2
		XGBoost	0.632 ± 0.006	0.546	0.687	858	4.5
		LightGBM	0.621 ± 0.007	0.399	0.683	6,105	5.0
S4	17	Random Forest	0.666 ± 0.009	0.467	0.726	3,208	27.7
		XGBoost	0.659 ± 0.006	0.562	0.715	869	5.0
		LightGBM	0.652 ± 0.004	0.378	0.727	6,139	5.1
S5	21	Random Forest	0.671 ± 0.005	0.446	0.732	3,204	25.2
		XGBoost	0.655 ± 0.005	0.504	0.716	878	4.3
		LightGBM	0.671 ± 0.010	0.415	0.736	6,145	6.7

These results support prioritising moisture and EC sensing, rather than a full soil-chemical panel, in any low-cost field-deployable monitoring system intended to flag elevated rhizome disease risk before symptoms become visible above ground. Sampling-optimisation approaches that couple interpretable machine learning with a deliberately reduced, high-value measurement set have shown that near-comparable predictive accuracy can often be retained while substantially cutting sensor and survey cost (Li et al., 2026), supporting the feasibility of a moisture-and-EC-only monitoring configuration for ginger. Mobile and Internet-of-Things-enabled decision support

systems have already demonstrated that sensor-derived predictions can be delivered directly to growers in real time for other high-value crops, whether through IoT-enabled yield forecasting with mobile decision support in strawberry production (Mumtaz et al., 2026) or through augmented-reality-enabled mobile crop recommendation and yield estimation tools (Kulkarni et al., 2026). Such platforms provide a plausible deployment pathway for the reduced-feature soil-disease model reported here, and a reduced sensor set would also lower the hardware cost and power demand of an edge-deployed monitoring device, a consideration directly relevant to subsequent resource-constrained implementation of this model. Because severity was highest in clay soils and wet, high-EC plots, such a device would be most useful where it triggers drainage checks and closer scouting in those plots.

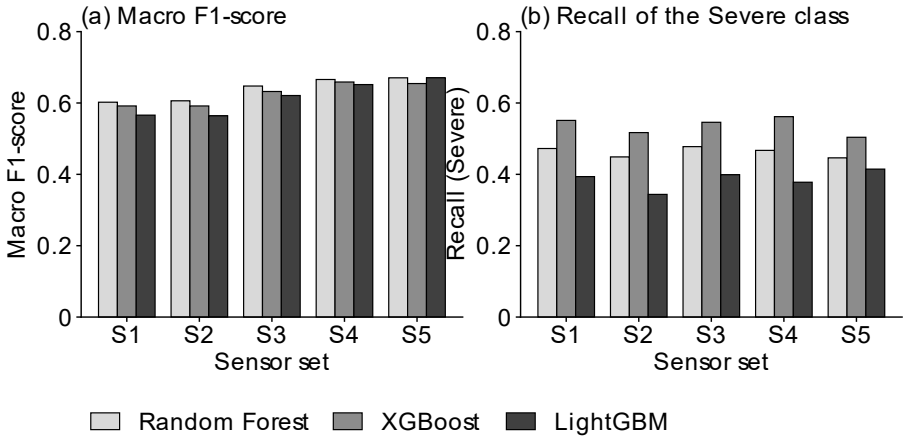

Figure 13. (a) Macro F1-score and (b) Severe-class recall of Random Forest, XGBoost and LightGBM for the five nested sensor sets S1–S5 (Table 12). Performance with moisture and EC alone (S1) is close to that of the full feature set (S5).

Limitations and future work

Several limitations should be noted. All data come from one district and one season, so the spatial and temporal tests reported above are internal checks, and validation on data from other regions and years remains the most important next step. The Severe class is small (127 samples) and severity was scored visually with criteria adapted from commonly used scales rather than a formal standard, so label noise at the Mild-Moderate and Severe boundary is likely and limits the attainable accuracy. Soil parameters alone cannot capture other determinants of rhizome rot, such as rainfall, microclimate, the health of the seed rhizome and field sanitation, which probably explain part of the 24–25% of test samples that were misclassified. In the imbalance benchmark, the SVM results depend on probability-based class labels, and the focal-loss objective is an approximation of the exact gradient. The size and latency figures are CPU measurements and not microcontroller tests. Future work will add weather and seed-rhizome variables, test the models on new seasons and regions, and deploy a compressed moisture-and-EC model on an edge device.

CONCLUSIONS

This study evaluated five machine learning algorithms for predicting three-class ginger rhizome disease severity from soil physicochemical parameters measured across 901 plots in Satara district, Maharashtra. The principal findings are as follows.

– Soil moisture content and electrical conductivity, individually and through their interaction with each other and with clay content, are the dominant soil-based predictors of rhizome disease severity in ginger. This was consistent across four modelling algorithms and in the SHAP analysis, in which high cumulative moisture, high EC and their product raised the predicted probability of Severe disease, and it is agronomically coherent with the water-dependent infection biology of *Pythium myriotylum*, the most frequently confirmed pathogen in the dataset.

– On the held-out test set, XGBoost gave the highest macro F1-score (0.696; accuracy 75.8%) and the highest F1-score for the Severe class (0.556), while Random Forest showed the most stable performance under group-wise cross-validation (macro F1-score 0.664). Differences between models were small relative to their uncertainty, and the test set contained only 16 Severe samples.

– Class weighting, SMOTE, ADASYN, Borderline-SMOTE and focal loss did not raise macro F1-score above that of the unweighted models. Oversampling increased Severe-class recall from 0.32–0.39 to 0.47–0.57 at the expense of precision, so the choice of strategy should reflect the relative cost of missed Severe plots and of false alarms.

– The Voting Ensemble achieved the highest macro ROC-AUC (0.884) but, together with LightGBM, the weakest Severe-class recall on the test set (0.375), illustrating that ensembling by probability averaging does not automatically improve minority-class detection. Rebalancing its members raised pooled Severe recall from 0.39 to 0.43–0.51 but did not improve its macro F1-score.

– Group-wise cross-validation, which respected the repeated-measures structure arising from repeated sampling of the same plots, produced a different model ranking to a single held-out test split, underscoring the importance of plot-aware validation when working with agricultural field data of this kind.

– Internal spatial and temporal validation gave macro F1-scores of 0.64–0.70 without a collapse in performance, and moisture and EC alone retained 84–90% of the macro F1-score of the full feature set. A three-feature XGBoost model built on these two sensors was below 1 MB, which supports the development of a minimal, low-cost soil-sensing module for resource-constrained (edge or TinyML) deployment for real-time rhizome disease risk monitoring in ginger fields.

– The dataset originates from a single agroclimatic region over one cropping season, and the Severe class remains comparatively small (127 of 1,141 samples). Validation on data from other regions and seasons is therefore needed before the models reported here are adopted operationally, and it will be the focus of our future work.

ACKNOWLEDGEMENTS**.** The authors thank the ginger growers of Satara district, Maharashtra, for access to their fields during soil sampling, and the technical staff involved in soil analysis and molecular pathogen confirmation. The authors gratefully acknowledge the support and resources provided by Shree Santkrupa Institute of Engineering & Technology, Ghogaon. This research was undertaken as part of a doctoral programme in Computer Applications under the Faculty of Science and Technology, Shivaji University, Kolhapur.

DECLARATION OF AI USE. The author(s) declare that artificial intelligence (AI) tools were used for structural organisation and preliminary drafting support. All AI-generated content was carefully reviewed, revised and validated by the author(s). The author(s) take full responsibility for the accuracy, integrity and originality of the final manuscript. No AI tools were used to generate data, fabricate references, or replace scholarly analysis.

REFERENCES

Asadollah, S.B.H.S., Jodar-Abellan, A. & Pardo, M.Á. 2026. SMART-Soil: Satellite-based machine learning approach for reliable tracking of soil moisture across CONUS. *Applied Soil Ecology* **225**. doi: 10.1016/j.apsoil.2026.107201

Asamoah, E., Heuvelink, G.B., Logah, V., Leenaars, J.G. & Bindraban, P.S. 2025. Fertilizer recommendations for maize production in Ghana: Comparison of machine learning, semi-mechanistic and conventional approaches. *European Journal of Agronomy* **174**. doi: 10.1016/j.eja.2025.127925

Bouyoucos, G.J. 1962. Hydrometer method improved for making particle size analyses of soils. *Agronomy Journal* **54**(5), 464–465.

Cao, Y., Yao, H., Zhang, C., Gu, C., Yang, L., Hao, R., Li, X., Wang, Z., Xiao, H., Wang, J. & Shi, Z. 2026. Data-driven assessment of nitrogen and phosphorus buffering capacity across 460 Chinese watersheds: Spatial patterns, drivers, and future projections. *Water Research* **293**. doi: 10.1016/j.watres.2026.125424

Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357. doi: 10.1613/jair.953

Chen, L., Hu, B., Sun, J., Xu, Y.J., Zhang, G., Ma, H. & Ren, J. 2025. Using remote sensing and machine learning to generate 100-cm soil moisture at 30-m resolution for the black soil region of China: Implication for agricultural water management. *Agricultural Water Management* **309**. doi: 10.1016/j.agwat.2025.109353

Chen, T. & Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. doi: 10.1145/2939672.2939785

Cui, S., Zhou, M., Yu, H., Xie, X., Xiao, L., Liu, J., Lin, J., Liu, X., Sui, Y. & Liu, J. 2025. Estimation of soil organic matter in mollisols based on artificial intelligence. *Ecological Informatics* **90**, 103223. doi: 10.1016/j.ecoinf.2025.103223

Deng, M., Zhang, C., Tang, M., Liao, C., Hu, Y., Zhang, Z., Feng, S. & Zheng, Z. 2025. Stacking ensemble learning coupled with multi-source remote sensing data: Enhancing soil salinity inversion accuracy in barley-cultivated salinized soils. *Agricultural Water Management* **322**, 109959. doi: 10.1016/j.agwat.2025.109959

Dong, Z., Wang, N., Xie, J. & Wang, T. 2025. Spectral data-driven and machine learning-based modeling of soil total nitrogen content. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **343**. doi: 10.1016/j.saa.2025.126583

Fares, A., Veettil, A.V., Rahman, A. & Awal, R. 2026. A review of artificial intelligence applications for predicting crop performance and enhancing food security under drought-induced water stress. *Agricultural Water Management* **325**, 110212. doi: 10.1016/j.agwat.2026.110212

Han, H., Wang, W.Y. & Mao, B.H. 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: *Advances in Intelligent Computing (ICIC 2005), Lecture Notes in Computer Science 3644*, pp. 878–887. doi: 10.1007/11538059_91

Hanway, J.J. & Heidel, H. 1952. Soil analysis methods as used in the Iowa State College Soil Testing Laboratory. *Iowa Agriculture* **57**, 1–31.

He, H., Bai, Y., Garcia, E.A. & Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks*, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969

He, J.H., Mickelson, S., Hatfield, J. & Tekeste, M.Z. 2025a. Tillage-induced soil feature extraction and multi-sensors fusion for tillage system classification. *Biosystems Engineering* **261**. doi: 10.1016/j.biosystemseng.2025.104329

He, S., Su, Y. & Guo, Z. 2025b. Shap-interpretable prediction of nutrient leaching risks in biochar-based slow-release fertilizers. *Science of the Total Environment* **1002**. doi: 10.1016/j.scitotenv.2025.180636

Huai, S., Zhang, Q., Jin, Y., Yu, W., Meersmans, J., Wang, S., Colinet, G. & Lu, C. 2025. Infrared spectroscopy for soil NPK estimation: Advances, challenges, and future directions in predictive modelling. *Trends in Analytical Chemistry* **185**. doi: 10.1016/j.trac.2025.118142

Jackson, M.L. 1973. *Soil chemical analysis*. Prentice Hall of India, New Delhi, 498 pp.

Karbasi, M., Randhawa, G.S., Farooque, A.A., Ali, M., Jamei, M., Khosravi, K., Afzal, H., Malik, A. & Zaman, Q.U. 2025. Utilizing soil characteristics and hybrid machine learning for interpretable potato yield prediction: A study with satin-bowerbird optimization and deep neural network. *Field Crops Research* **338**. doi: 10.1016/j.fcr.2025.110311

Katipoğlu, O.M., Kartal, V., Akiner, M.E., Terzioğlu, Z.Ö. & Kılıç, Z. 2025. Prediction of soil moisture via feature selection, model optimization, and climate data integration. *Physics and Chemistry of the Earth* **142**. doi: 10.1016/j.pce.2025.104250

Kaur, A., Malekian, R., Randhawa, G.S., Farooque, A.A., Garmdareh, E.H., Acharya, B., Singh, R. & Selopal, G.S. 2026. Advancing sustainable agriculture in Atlantic Canada through HYDRA-SE: A hybrid decision and regression stacked ensemble for yield prediction. *Advanced Engineering Informatics* **71**. doi: 10.1016/j.aei.2026.104433

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* **30**, 3146–3154.

Khosravi Aqdam, K., Asadzadeh, F., Rezapour, S., Nouri, A. & Shabani, F. 2026. An integrated soil health and machine learning framework for quantifying soil degradation in semi-arid agricultural lands. *Soil & Tillage Research* **259**. doi: 10.1016/j.still.2026.107099

Kulkarni, R.R., Deshpande, V.V., Ramkumar Menon, A. & Eswaran, M. 2026. An augmented reality enabled mobile decision support system for intelligent crop recommendation and yield estimation. *Results in Engineering* **30**. doi: 10.1016/j.rineng.2026.110853

Lei, Y., Kong, J., Li, X., Han, B., Lu, M. & Song, Y. 2026. Spatial mapping of soil profile pH in coastal saline-alkali farmland: A novel approach integrating stacking ensemble model with spatial interpretability analyses. *Environmental Technology & Innovation* **43**. doi: 10.1016/j.eti.2026.105078

Lemaître, G., Nogueira, F. & Aridas, C.K. 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* **18**(17), 1–5.

Li, S., Feng, W., Yi, X., Shan, X., Xue, Z., Guo, C., Tang, X. & Zhao, J. 2026. Sampling optimization and interpretable machine learning for high-resolution soil thickness mapping in mountainous terrain. *Engineering Applications of Artificial Intelligence* **177**. doi: 10.1016/j.engappai.2026.114898

Li, Z., Cheng, Q., Chen, Z., Xiang, Y., Hu, X., Lazarovitch, N. & Zhen, J. 2025. Estimating soil water content of cotton fields using UAV-based multi-source remote sensing data fusion. *Agricultural Water Management* **322**, 109996. doi: 10.1016/j.agwat.2025.109996

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988. doi: 10.1109/ICCV.2017.324

Lundberg, S.M. & Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**, 4765–4774.

Mumtaz, A., Mohd Noor, A.S., Khan, M.R., Bano, A., Muhaimin, A., Mohamad, M.S. & Che Yahya, C.A. 2026. IoT-Enabled Machine Learning Framework for Strawberry Yield Prediction with Mobile Decision Support. *Procedia Computer Science* **280**, 606-613. doi: 10.1016/j.procs.2026.04.077

Nautiyal, M., Joshi, S., Hussain, I., Rawat, H., Joshi, A., Saini, A., Kapoor, R., Verma, H., Nautiyal, A., Chikara, A., Ahmad, W. & Kumar, S. 2025. Revolutionizing agriculture: A comprehensive review on artificial intelligence applications in enhancing properties of agricultural produce. *Food Chemistry: X* **29**, 102748. doi: 10.1016/j.fochx.2025.102748

Ngu, N.H., Chotpantarat, S., Nhi, P.T.P., An, N.T.T. & Thanh, N.N. 2025. Effects of agricultural land-use types on soil physicochemical parameters in Central Vietnam using machine learning. *Applied Soil Ecology* **217**. doi: 10.1016/j.apsoil.2025.106621

Olsen, S.R., Cole, C.V., Watanabe, F.S. & Dean, L.A. 1954. *Estimation of available phosphorus in soils by extraction with sodium bicarbonate*. USDA Circular 939. U.S. Government Printing Office, Washington, DC, 19 pp.

P, H., Gautam, S., Joshi, S.K., Asirvatham, L.G. & Rakshith, B.L. 2025. Advancements in soil moisture estimation through integration of remote sensing and artificial intelligence techniques. *Science of the Total Environment* **1001**. doi: 10.1016/j.scitotenv.2025.180503

Palta, P., Palta, A. & Kumar, A. 2025. Fertilizer-induced variations in soil dielectric properties: Implications for smart agriculture and predictive modelling. *Physics and Chemistry of the Earth* **142**. doi: 10.1016/j.pce.2025.104240

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Raja Ranganathan, S., Alzaben, N., Alghamdi, A.G. & Dafaalla, H. 2026. Cloud-based soil salinity mapping using remote sensing and machine learning approaches. *Journal of African Earth Sciences* **236**. doi: 10.1016/j.jafrearsci.2026.105993

Ribeiro, M.T., Singh, S. & Guestrin, C. 2016. 'Why should I trust you?': Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. doi: 10.1145/2939672.2939778

Schweng, S., Bernardini, L., Keiblinger, K., Kaul, H.P., Jr., I.F., Lukač, N., Del Ser, J. & Holzinger, A. 2025. What can artificial intelligence do for soil health in agriculture?. *Computer Science Review* **59**. doi: 10.1016/j.cosrev.2025.100832

Song, J. & Yi, Y. 2025. Cascaded Machine Learning of Soil Moisture and Salinity Prediction in Estuarine Wetlands Based on In Situ Internet of Things Monitoring. *Water Resources Research* **61**(4). doi: 10.1029/2024WR038271

Subbiah, B.V. & Asija, G.L. 1956. A rapid procedure for the estimation of available nitrogen in soils. *Current Science* **25**, 259–260.

Vahidi, M., Shafian, S. & Frame, W.H. 2025. Multi-depth soil moisture estimation via 1D convolutional neural networks from drone-mounted ground penetrating Radar data. *Computers and Electronics in Agriculture* **232**. doi: 10.1016/j.compag.2025.110104

Walkley, A. & Black, I.A. 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Science* **37**(1), 29–38. doi: 10.1097/00010694-193401000-00003

Wang, H., Safiul Azam, F.M., Xiao, X., Tasnim, F., Das, P., Zhou, X. & Gan, H. 2025a. Design and optimization of NPK fertilization for ginger (Zingiber officinale) seed rhizome production from tissue-cultured plantlets in soilless culture. *Scientia Horticulturae* **348**. doi: 10.1016/j.scienta.2025.114208

Wang, J., He, P., Du, K., Li, X., Xu, L., Yang, F. & Bi, R. 2025b. Machine learning-based inversion and sensitivity analysis of soil moisture in Hemerocallis cultivation. *Agricultural Water Management* **322**. doi: 10.1016/j.agwat.2025.110011

Wang, Y., Yang, P., Zhou, L., Wang, H. & Li, Z. 2026. Machine learning–based spatiotemporal mapping and risk zoning of soil salinization in an arid inland river basin: A case study of the Weigan River Basin, northwestern China. *Agricultural Water Management* **330**. doi: 10.1016/j.agwat.2026.110412

White, T.J., Bruns, T., Lee, S. & Taylor, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M.A., Gelfand, D.H., Sninsky, J.J. & White, T.J. (eds) *PCR Protocols: A Guide to Methods and Applications*. Academic Press, San Diego, pp. 315–322.

Xu, B., Yang, J., Zhao, R., Li, J., Ning, X., Li, F., Lin, H., Yang, Z. & Fan, L. 2026. Multi-dimensional coupling framework for soil salinity assessment: Revealing spatiotemporal interactions among soil, environment, and cotton growth. *Agricultural Systems* **238**. doi: 10.1016/j.agsy.2026.104853

Yang, M.D., Su, T.C. & Hsu, Y.C. 2026. Estimating fine-scale soil moisture content and wheat yield in drought-heterogeneous farmlands using machine learning regressions on multispectral UAV imagery. *Computers and Electronics in Agriculture* **253**. doi: 10.1016/j.compag.2026.112145

Zhai, Y., Zhao, W., Dong, K., Wang, Y., Lin, H., Dong, B., Wang, D., Xu, Y. & Liu, J. 2026. Explainable machine learning reveals key determinants of biochar soil-application effects on plant growth. *Environmental Technology & Innovation* **41**. doi: 10.1016/j.eti.2026.104829

APPENDIX

Table A1 lists the hyperparameter search ranges and the values selected for the models on the full three-class dataset. All other hyperparameters were left at library defaults except for the fixed settings listed below the table; the regularisation parameters of XGBoost (L2 $\lambda = 1$, L1 $\alpha = 0$, $\gamma = 0$, minimum child weight = 1) and LightGBM (L1 = L2 = 0, minimum split gain = 0) were therefore not tuned. Table A2 gives the computing environment.

Table A1. Hyperparameter search ranges and selected values (randomised search, scoring by macro F1-score, three-fold GroupKFold on the training partition, random seed 42)

Model	Hyperparameter	Search range	Selected
SVM	C	[0.1, 0.3, 1, 3, 10, 30, 100]	3
	gamma	[scale, auto, 0.001, 0.01, 0.1]	auto
	kernel	[rbf]	rbf
Random Forest	n_estimators	[200, 300, 400, 500]	200
	max_depth	[4, 6, 8, 12, None]	12
	min_samples_leaf	[1, 2, 4, 8]	4
	max_features	[sqrt, log2, 0.5]	log2
XGBoost	n_estimators	[100, 150, 200, 300]	200
	max_depth	[3, 4, 5]	4
	learning_rate	[0.03, 0.05, 0.1]	0.03
	subsample	[0.8, 1.0]	0.8
	colsample_bytree	[0.7, 1.0]	0.7
LightGBM	n_estimators	[100, 150, 200, 300]	300
	max_depth	[-1, 4, 6, 8]	-1
	learning_rate	[0.03, 0.05, 0.1]	0.03
	num_leaves	[15, 31, 63]	63
	subsample	[0.8, 1.0]	1.0
	min_child_samples	[5, 10, 20]	5

Candidates evaluated: 15 (SVM, Random Forest) and 10 (XGBoost, LightGBM). Fixed settings: SVM, balanced class weights and probability estimates enabled for the final model; Random Forest, balanced class weights and Gini criterion with bootstrap sampling; XGBoost, multi:softprob objective, mlogloss metric and balanced sample weights; LightGBM, multiclass objective and balanced class weights; Voting Ensemble, soft voting of the four tuned models; random seed 42 throughout.

Table A2. Computing environment used for model training and evaluation

Item	Specification
Platform	Google Colab runtime, Linux, 2 CPU cores
Python	3.13.15
scikit-learn	1.6.1
NumPy / pandas	2.1.3 / 2.2.3
XGBoost	3.4.1
LightGBM	4.6.0
imbalanced-learn	0.14.2
SHAP	0.52.0
Random seed	42
Search	RandomizedSearchCV, macro F1-score, three-fold GroupKFold
Imbalance benchmark	Five-fold StratifiedGroupKFold, three repetitions (seeds 42–44)