International Journal of Research in Modern Engineering & Emerging Technology (IJRMEET)
Vol. 14, Issue: 10, October 2026
(IJRMEET) ISSN (o): 2320-6586

Explainable Prediction of Crop Disease Risk Using IoT Based Environmental Data: A Field Oriented Machine Learning Framework for Chhattisgarh

DOI: https://doi.org/10.63345/ijrmeet.org.v14.i10.1

Preeti Mishra
Research Scholar, Department of Computer Science
Kalinga University, Naya Raipur, Chhattisgarh
Dr. Anupa Sinha
Assistant Professor, Department of Computer Science
Kalinga University, Naya Raipur, Chhattisgarh

Abstract— Environmental sensors can indicate conditions associated with crop disease, but a sensor reading alone cannot confirm an infection. This methodological paper develops an explainable machine learning framework for assessing rice disease risk in Chhattisgarh using time-stamped Internet of Things (IoT) observations. It defines field–date observations, a seven-day warning horizon, expert-confirmed outcomes and predictors derived only from earlier measurements. Random Forest, a support vector machine and a seasonal baseline are specified for comparison. The evaluation design separates fields and later periods during validation and examines discrimination, probability calibration, alert burden and model explanations. The contribution is a reproducible protocol for evaluating a field warning system; no new Chhattisgarh field dataset or verified predictive performance is claimed.

Keywords: Chhattisgarh; rice disease; Internet of Things; Random Forest; explainable artificial intelligence; probability calibration; field validation.

Introduction

Chhattisgarh is a relevant setting for rice focused decision support because rice occupies a substantial part of its cultivated area and many fields are rainfed [1]. Temperature, moisture and canopy wetness can change the conditions under which disease develops. A sensor, however, measures environmental exposure rather than the presence of a pathogen. A risk estimate is consequently a prompt to inspect the crop, not a diagnosis and not a treatment instruction.

Automated plant disease research has demonstrated the potential of image classification, although models trained on controlled images may perform differently under field conditions [2]. This paper develops a complementary framework that uses antecedent environmental observations to estimate the risk of a subsequently confirmed rice disease event in Chhattisgarh. It specifies the outcomes, predictors, validation design and explanations needed to assess such a system under local growing conditions.

Research Problem and Contribution

The research question is whether time-stamped environmental measurements can predict an expert-confirmed rice disease event in a previously unseen field or later growing period with useful discrimination, calibrated probabilities and intelligible explanations. A random split of individual sensor records may overstate performance if records from the same field or disease episode appear in both training and test sets.

The methodological contribution combines antecedent environmental windows, expert-defined outcomes, field- and time-separated validation, and an audit of both global and individual explanations. The study design is intended for prospective evaluation; field observations from Chhattisgarh have not been provided for analysis in this paper.

Review of Related Work

Mohanty, Hughes and Salathé [2] established a widely cited image based baseline using the PlantVillage dataset. Image symptoms are valuable after appearance, whereas environmental histories can be examined before an expert confirms disease. These data types answer different questions and should only be fused when records refer to the same field, date and crop stage. Lundberg and Lee [3] introduced SHAP values for allocating a model prediction among input features. Such explanations can describe the model's computation; they cannot establish that humidity or any other feature caused an outbreak.

Repeated measurements from the same farm are statistically dependent. Group based validation prevents a field from appearing in training and test folds [4]. For warning systems, class recall alone is insufficient: predicted probabilities should also be checked with reliability plots

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

and proper scores [5]. Permutation importance provides a held out measure of predictive dependence on a feature, although correlated environmental variables can complicate its interpretation [6]. These principles address shortcomings of a simple random split and an unqualified list of Random Forest feature importances.

Objectives

1. To specify an auditable field-level dataset for rice disease risk in selected districts of Chhattisgarh.
2. To compare a seasonal baseline, Random Forest and a support vector machine on identical held-out observations when a labelled dataset becomes available.
3. To define measures of discrimination, calibration, operational burden and explanation stability across fields and time periods.

Research Questions

1. Does sensor history improve prediction of a verified disease event within a fixed warning horizon over a seasonal baseline?
2. How much performance is lost when evaluation moves from random records to unseen fields and subsequent time periods?
3. Are the conditions highlighted in individual alerts stable, and do experts consider the explanations agronomically plausible?

Materials and Methods

- **Study setting and unit of analysis.** The proposed observation unit is a field–date pair, indexed by field identifier f and date t. Participating rice-growing districts should represent relevant agroecological variation and feasible field access. For each field, the dataset should include district, privacy-protected location, rice variety, transplanting or sowing date, crop stage and dated management events. The disease target, recruitment period, fields and seasons should be specified before model fitting; no achieved sample size is reported here.

- **Outcome definition and ground truth.** For a prespecified disease, Y(f,t)=1 if an agronomist confirms a new event within seven days after date t. Y(f,t)=0 only when a field assessment during that interval finds no event. Unobserved intervals should not be labelled negative. The disease definition, symptom criteria, assessment date, assessor, severity and supporting image should be recorded, with disagreements resolved through a second assessment. The seven-day horizon is an illustrative protocol choice requiring agronomic review. Pest events require a separate endpoint.

- **Sensor and contextual variables**

Variable	Collection and processing	Interp
Air temperature	DHT22 or calibrated equivalent; hourly summaries	Candidate microcl
Relative humidity	Hourly and rolling window summaries	Moisture related e
Soil moisture	Capacitive sensor; site calibration and units	Water status proxy
Leaf wetness	Validated leaf wetness sensor or specified proxy	Canopy wetness d
Rainfall	Gauge or attributable station data, with provenance	Antecedent rain ar
Crop stage and management	Field log with date of entry	Potential confound

Fog-related readings and measured leaf wetness represent different variables. A fog sensor should not be treated as a leaf-wetness instrument unless its output has been validated against a suitable reference. A field dataset should document sensor placement, sampling intervals, calibration, power interruptions and missing readings. If leaf images are subsequently included, their labels, collection procedures and model evaluation should be documented separately.

- **Preprocessing and feature construction.** Sensor readings should be converted to a common time zone and an hourly grid. Implausible readings should be flagged against instrument specifications while raw data are retained. Predictors may include daily summaries, cumulative wetness hours and rainfall during the preceding 24, 72 and 168 hours, using information available no later than date t. Imputation and scaling must be learned from training data within each validation fold. Readings from different fields should never be forward-filled across field boundaries; future observations and post-diagnosis treatments must not enter the predictors.

- **Models and training.** The protocol compares a seasonal event-prevalence baseline, a Random Forest classifier and a support vector machine with a radial basis function kernel. Hyperparameters should be selected using grouped folds within the training partition. Class weighting, probability calibration and alert thresholds should be fitted without access to the final test data. The modelling report should specify software versions, random

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

seeds and a de-identified data dictionary where sharing is permitted.

- **Validation design and leakage control.** A later growing season should provide a chronological test set where available. Within the remaining records, entire fields should be reserved for validation, with grouped cross-validation used for model selection [4]. Records from the same field or disease episode must not cross partition boundaries. If only one season is available, a contiguous later-period test is a weaker temporal assessment and should be labelled as such. Probability calibration should use validation data separate from final testing. District-level sensitivity analysis is appropriate only when enough districts and disease events are represented.
- **Explainability and advisory rules.** Held-out permutation importance may identify variables on which a fitted model depends [6]. Local SHAP values may describe the variables contributing to an individual model output [3]. Each displayed alert should include its date, predicted probability, data completeness and a plain-language explanation. An attribution describes model behaviour rather than a cause of disease. A threshold chosen on validation data may prompt field inspection; diagnosis and treatment require expert confirmation and current local agronomic advice.

Analytical Approach

The evaluation framework uses environmental variables from an IoT monitoring system and compares Random Forest with a support vector machine and a seasonal baseline. Classification accuracy is included for completeness, while event recall, precision and false-alert burden are more informative for assessing a warning system. These are specified evaluation measures, not results calculated from a Chhattisgarh dataset.

The primary proposed operational measure is event recall at a fixed inspection workload, expressed as false alerts per 100 field-days. Secondary measures include precision, specificity, F1 score, the area under the precision–recall curve, the area under the receiver operating characteristic curve and probability calibration. Recall captures missed disease events; precision captures the proportion of alerts linked to verified events. Their numerical values require held-out predictions and verified outcomes.

The measures are defined as follows:

- **Precision** = TP/(TP + FP)
- **Recall** = TP/(TP + FN)
- **Specificity** = TN/(TN + FP)
- **F1 score** = 2 × Precision × Recall/(Precision + Recall)

TP denotes true positives, FP false positives, FN false negatives and TN true negatives. These counts must come from a clearly identified held-out test set. An accuracy percentage by itself cannot recover the confusion matrix or establish event recall and precision.

Calibration should be examined using a reliability plot, calibration slope and intercept, and the Brier score, defined for binary outcomes as $(1/N)\Sigma(p_i - y_i)^2$, where p_i is predicted risk and y_i is the observed outcome. Such statistics cannot be calculated without record-level probabilities and verified outcomes. Uncertainty intervals should resample entire fields, with seasonal blocking if sufficiently many seasons are observed.

- **Comparative Analysis**

Analysis	Approach described or examined	Interpretation
Model discrimination	Seasonal baseline, Random Forest and SVM on the same held-out records	Compare recall, precision and discrimination; no ranking is claimed without results.
Field generalization	Field-separated and later-period test sets	Estimate transfer to unseen fields and subsequent conditions.
Calibration	Reliability plot, Brier score, calibration slope and intercept	Assess whether predicted probabilities reflect observed event frequencies.
Data contribution	Weather-only, local-sensor-only and combined inputs	Determine whether local sensors add predictive information.
Operational value	Event recall at a fixed number of false alerts per 100 field-days	Balance missed events against field inspection effort.
Explainability	Held-out permutation	Examine attribution stability and agronomic

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Analysis	Approach described or examined	Interpretation
	importance and local SHAP analysis	plausibility.

Analysis and Interpretation of the Framework — The proposed pipeline connects time-stamped environmental readings with a future expert-confirmed disease event at the field–date level. A Random Forest can accommodate nonlinear associations among environmental variables; an SVM and a seasonal baseline provide comparison methods. A web interface can present a model score and its explanation once a model has undergone field validation. These architectural and analytic components define how the research question would be tested; they do not constitute measured predictive results.

A numerical comparison of the candidate models requires a labelled dataset, a fixed test partition and a confusion matrix. These records were not available for this paper. No accuracy estimate, class-wise performance figure or model ranking is therefore presented as a Chhattisgarh field result.

Component	Study description	Interpretation
Observation unit	Field–date record with historical sensor readings	Keeps inputs prior to the outcome and identifies repeated field observations.
Outcome	Expert-confirmed disease event within a defined horizon	Separates measured conditions from actual disease diagnosis.
Candidate models	Seasonal baseline, Random Forest and SVM	Enables matched comparisons once labelled data exist.
Evaluation	Field-separated and later-period tests	Assesses transfer beyond training fields and times.
Explanation	Global and local model attribution	Describes model behaviour; causal interpretation is not implied.

The analysis defines how environmental observations can be converted into risk predictors and how a Random Forest could be evaluated against alternatives. It does not establish sensitivity to disease outbreaks, the reliability of predicted probabilities or a reduction in crop loss. Those claims require expert-confirmed field events and prospective assessment.

Discussion

The framework brings environmental monitoring and machine learning together at the level of a dated field observation. Sensor history may reveal conditions preceding visible disease, making a risk score useful for prioritising inspection. The score cannot confirm infection or justify treatment without field verification.

Repeated measurements from the same field are dependent. When records from one field or disease episode are divided across training and test sets, an evaluation can overstate performance on unfamiliar fields. The proposed field-separated and later-period tests address this concern. Their outcomes cannot be stated until labelled field data are analysed.

Interpretability is essential for examining individual alerts. Global importance can identify variables used by a model, while local explanations can identify those associated with a specific score. Neither method demonstrates that a variable caused the disease. Correlation among humidity, rainfall and wetness can also affect the apparent contribution of each variable; explanation stability should therefore be assessed.

Field implementation would require sensor calibration, complete time records, power and connectivity monitoring, and expert-confirmed labels. Observations should be stored locally during network interruptions for later transmission. Farmer identity and precise coordinates should be protected through consent and restricted access.

Limitations and Scope for Further Validation

The paper sets out a predictive framework and IoT monitoring architecture. It does not analyse an auditable Chhattisgarh field dataset and cannot report the number of participating fields, expert-confirmed disease events, class-wise performance, calibration or alert lead time. No reduction in pesticide use or crop loss is claimed.

Prospective validation should target a clearly identified rice disease and a defined warning horizon. Time-stamped sensor observations should be linked to independent field assessments, and testing should use fields and later periods excluded from model training. Reporting event counts, confusion matrices, uncertainty intervals, calibration and false-alert burden would make performance assessable and reproducible.

A later study could evaluate paired field images and environmental histories. It should compare a combined

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

model with image-only and sensor-only baselines on the same held-out fields to determine whether fusion improves performance.

Conclusion

This paper specifies an explainable IoT-supported framework for rice disease risk assessment in Chhattisgarh. It defines field-level disease outcomes, antecedent environmental predictors, candidate machine learning models, leakage-resistant validation and explanations for individual alerts. These elements form a testable research protocol rather than a claim of established local predictive accuracy.

The framework can be evaluated when expert-labelled local observations become available. Independent field and temporal tests, classification measures, probability calibration and inspection workload are needed before its predictions can be treated as dependable in practice.

References

[1] Indian Council of Agricultural Research. "Chhattisgarh," state agricultural profile. https://www.icar.gov.in/node/17261 (accessed 24 September 2026).

[2] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," Frontiers in Plant Science, vol. 7, article 1419, 2016. doi:10.3389/fpls.2016.01419.

[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, 2017. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[4] scikit-learn developers, "Cross-validation: evaluating estimator performance," documentation, section on grouped cross-validation. https://scikit-learn.org/stable/modules/cross_validation.html (accessed 24 September 2026).

[5] scikit-learn developers, "Probability calibration," documentation. https://scikit-learn.org/stable/modules/calibration.html (accessed 24 September 2026).

[6] scikit-learn developers, "Permutation feature importance," documentation. https://scikit-learn.org/stable/modules/permutation_importance.html (accessed 24 September 2026).

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).