Beyond Accuracy: A Bayesian Metric for Binary Classifier Evaluation in Imbalanced Crop Disease Data

Aquiles Darghan*, Ivan Lizarazo, Carlos Rivera, Liliana Castillo-Villamor*, and Jorge Jola

Departamento de Agronomía, Facultad de Ciencias Agrarias, Universidad Nacional de Colombia, Bogotá, Colombia
Email: aqedarghanco@unal.edu.co (A.D.); ializarazos@unal.edu.co (I.L.); caariveramo@unal.edu.co (C.R.); lcastillov@unal.edu.co (L.C.V.); jjola@unal.edu.co (J.J.)
*Corresponding author

Manuscript received April 1, 2026; accepted June 3, 2026; published July 14, 2026

***Abstract*—When classes are imbalanced, the evaluation of binary classification algorithms presents a structural problem: the prevalence of the positive class varies widely across practical applications, from fraud detection and system monitoring to disease identification in medical and epidemiological settings. In these scenarios, classical accuracy yields misleadingly high values that fail to capture true discriminative capacity, as it is operates as a linear function of prevalence, dominated by the majority class and unable to penalize the trivial classifier that labels everything as the dominant class. To address this limitation, Bayesian Geometric Accuracy, defined as the geometric mean of the positive and negative predictive values, is proposed. Unlike classical accuracy, this metric is nonlinear in prevalence, vanishes at both extremes, and penalizes class imbalance by construction because predictive values explicitly depend on prevalence through Bayes' theorem. The equivalence condition between both measures yields a degree 3 polynomial in prevalence, whose coefficients are expressed in closed form in terms of sensitivity, specificity, and the Youden index. This polynomial has exactly two roots in the unit interval, and their separation constitutes a principled measure of the classifier's discriminative strength. The framework is illustrated through a multilayer perceptron trained on spectral indices for the binary classification of disease presence in potato crops, where classical accuracy is 0.868 while Bayesian Geometric Accuracy yields 0.655, a gap that reveals a substantial overestimation at the observed prevalence of 12.7%. The proposed metric defines a principled safe interval of class ratios that may guide a more robust model design for detecting early crop disease.**

***Keywords*—Bayesian accuracy, positive predictive value, negative predictive value, class imbalance**

I. Introduction

In supervised learning, reporting model accuracy has become a routine proxy for performance. In binary classification, the evaluation relies on the following metrics derived from the confusion matrix: sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) [1]. In the presence of class imbalance, these metrics can be misleading [2], which is why tabular summaries of multiple metrics have become standard practice in well-established libraries, such as yardstick [3] and caret [4] packages in R, and scikit-learn [5] and XGBoost [6] in Python. Despite this wide availability of metrics, one particularly relevant limitation remains unappreciated: the standard metrics are computed conditional on a specific data partition, which implicitly treats the observed class distribution as fixed and representative of the target population.

Classical accuracy is a linear combination of sensitivity and specificity weighted by prevalence. When the distribution of classes in the test set does not reflect real-world prevalence at deployment, the reported accuracy fails to describe the probability that a given prediction is correct. Instead of treating prevalence as an accidental result of data partitioning, it is proposed to incorporate it as prior information in the evaluation process [7].

This Bayesian reformulation shifts the perspective from truth-conditioned performance to prediction-conditioned reliability. Classical sensitivity and specificity quantify accuracy conditional on the true class, whereas PPV and NPV assess correctness conditional on the predicted class. Accordingly, we define and interpret $B_{se} = PPV$ (Bayesian sensitivity) and $B_{sp} = NPV$ (Bayesian specificity) as the classical intrinsic metrics' prediction-space counterparts [8].

The Bayesian accuracy proposed here reorients the assessment around a deployment-centered statement: a prediction is correct if its associated predictive value confirms it. We propose the geometric mean of Bayesian sensitivity and specificity rather than arithmetic weighting of predictive values. Just as the Youden index combines classical sensitivity and specificity through their sum, the geometric mean combines their Bayesian counterparts in a form that is simultaneously more conservative, penalizing imbalance between predictive values more aggressively than any arithmetic average. Furthermore, it exhibits greater sensitivity to extreme prevalence regimes, vanishing whenever the predictive value approaches zero. A metric anchored in the prediction space is the conceptually appropriate response to a field where classical accuracy metrics have consistently failed to reflect the utility of real-world classifiers [9].

This study focuses on expanding the understanding of the limitations of classical metrics as they apply to accuracy and its derived metrics [10]. Section II presents the binary classification framework by introducing the confusion matrix, deriving classical intrinsic metrics, developing their predictive counterparts (Bayesian sensitivity and specificity), and formally defining the proposed Bayesian geometric accuracy. Section III establishes the algebraic structure of the equivalence condition between classical and Bayesian accuracy, which yields a degree3 polynomial in prevalence. Section IV explores the analytical behavior of critical prevalence, defines the interior intervals, and details the theoretical connections with both the classical Youden index and its Bayesian counterpart. Section V illustrates the full framework through a Multilayer Perceptron (MLP) model trained on spectral indices for binary classification of disease

presence in potato crops in a sample of $n = 212$ observations.

II. Sensitivity and Specificity of Binary Classification

A. Confusion Matrix

The confusion matrix is the starting point for evaluating any binary classification: a 2×2 contingency table contrasting observed class labels y against algorithmic predictions $\hat{y}$. Each observation falls into one of four cells: true positives (*TP*), false negatives (*FN*), false positives (*FP*), and true negatives (*TN*) [11]. The observed classes index the rows (O_+, O_-), and the algorithmic predictions index the columns (C_+, C_-), as shown in Table 1.

Table 1. Confusion matrix for a binary classifier. Rows: observed class (O_+ = positive, O_- = negative). Columns: algorithmic prediction ($\hat{y} = C_+$ predicted positive, $\hat{y} = C_-$ predicted negative)

True condition	**Value classified by the algorithm**		**Sub-total**
Class label	$\hat{y} = C_+$	$\hat{y} = C_-$	
Observed value $y = O_+$	$n(TP)$	$n(FN)$	$n(O_+)$
Observed value $y = O_-$	$n(FP)$	$n(TN)$	$n(O_-)$
Subtotal	$n(C_+)$	$n(C_-)$	n

The row marginals $n(O_+)$ and $n(O_-)$ are the total truly positive and truly negative cases, respectively. The column marginals $n(C_+)$ and $n(C_-)$ are the total positive and negative predictions. The prevalence of the positive class is $P = n(O_+)/n$, where n is the total number of observations, such that $1-P$ denotes the negative class prevalence. This single quantity governs the behavior of all the Bayesian metrics developed in the following sections [12].

B. Classical Metrics: Sensitivity, Specificity, and the Youden Index

The sensitivity (C_{se}) is the probability that the algorithm classifies a truly positive case as positive: $C_{se} = p(O_+)$ and the specificity (C_{sp}) is the probability of correctly classifying a truly negative case: $C_{sp} = p(O_-)$.

This prevalence independence is not an empirical claim about classifier behavior across different deployment contexts —it is the defining formal property of intrinsic metrics as opposed to predictive metrics, and it holds when the classifier's decision rule is fixed. In practice, a classifier retrained on a rebalanced dataset will generally produce different values of C_{se} and C_{sp}, because the decision boundary shifts in response to the altered class distribution. The present framework explicitly separates these two levels: the theoretical analysis in Sections III and IV operates conditional on fixed intrinsic parameters, characterizing how the metric landscape changes with prevalence for a given classifier; the discussion of rebalancing in Section VI addresses the separate empirical question of what happens when the classifier itself is modified. Conflating these two levels leads to the mistaken conclusion that invariance is assumed about the real world rather than about the fixed-classifier evaluation problem. Nevertheless, fixing the classifier does not eliminate the deployment dilemma: given a positive prediction within a specific prevalence context, the probability that it is correct depends entirely on the condition´s prevalence in the population being tested.

The classical Youden index (C_J) combines both metrics into a single discriminative summary:

$$C_J = C_{se} + C_{sp} - 1. \tag{1}$$

It ranges from −1 (total misclassification) to 1 (perfect discrimination), with $C_J = 0$ for a random classifier. Because both C_{se} and C_{sp} are prevalence independent, C_J is a fixed horizontal reference in any classifier performance against *P*. This fixed benchmark is central to Section IV, where Bayesian metrics are defined and assessed precisely through their relationship to C_J. Finally, the classical accuracy is

$$C_A = PC_{se} + (1-P)C_{sp}. \tag{2}$$

This is a weighted average of sensitivity and specificity with the prevalences of the classes as weights and is linear in *P*. A fundamental identity connecting C_A to the prediction space is derived in the next section [13].

C. Bayesian Counterparts: Predictive Values

Applying Bayes's theorem and conditioning on the prediction rather than the true class yields two complementary metrics. The positive predictive value (*PPV*) is the probability that a predicted positive is truly positive:

$$B_{se} = PPV = p(C_+) \tag{3}$$

The negative predictive value (*NPV*) is the probability that a predicted negative is truly negative:

$$B_{sp} = NPV = p(C_-) \tag{4}$$

Expressed as explicit functions of the prevalence *P* through Bayes' theorem,

$$B_{se} = \frac{PC_{se}}{PC_{se}+(1-P)(1-C_{sp})}. \tag{5}$$

$$B_{sp} = \frac{(1-P)C_{sp}}{(1-P)C_{sp}+P(1-C_{se})}. \tag{6}$$

Both are rational functions of P in (0,1) that vary continuously. Their boundary behavior is: at $P \to 0$, $PPV \to 0$ $NPV \to 1$; at $P \to 1$, $PPV \to 1$ and $NPV \to 0$. This correct behavior at prevalence extremes is a key property that the proposed Bayesian geometric index (B_G) inherits, and that the naive Bayesian accuracy does not (Section III.A).

The marginal prediction probabilities are:

$$A = p(\hat{y} = C_+) = PC_{se} + (1-P)\big(1 - C_{sp}\big) \tag{7}$$

$$B = p(\hat{y} = C_-) = P(1 - C_{se}) + (1-P)C_{sp} \tag{8}$$

A fundamental identity (with $A + B = 1$) follows directly from Eqs. (3–8):

$$AB_{se} + (1-A)B_{sp} = C_A. \tag{9}$$

D. Definition of the Bayesian Geometric Index B_G

The appropriate summary of joint predictive performance must satisfy three minimum requirements: (i) vanish when either predictive value is zero (correct boundary behavior);

(ii) be symmetric under re labeling of classes (invariance); and (iii) penalize asymmetry between *PPV* and *NPV* (conservatism) [14, 15]. The geometric mean satisfies all three conditions.

Definition (Bayesian geometric accuracy). *For a* classifier *with sensitivity C_{se} and specificity C_{sp} in (0,1), the Bayesian geometric accuracy is defined for $P \in (0,1)$ as:*

$$B_G = \sqrt{B_{se}B_{sp}} \tag{10}$$

In terms of the confusion matrix cell counts and using the marginal prediction probabilities A and B:

$$B_G = \sqrt{\frac{P(1-P)C_{se}C_{sp}}{AB}} \tag{11}$$

1) Analytical properties

Property 1 —Correct boundary behavior

$B_G \to 0$ as $P \to 0$ and as $P \to 1$, correctly reflecting that the joint predictive utility is zero when the positive class is absent or universal. This follows directly from the boundary behavior of *PPV* and *NPV*: as $P \to 0$, $PPV \to 0$ while $NPV \to 1$ and as $P \to 1$, $NPV \to 0$ while $PPV \to 1$; thus, in both extremes one factor of the geometric mean vanishes and pulls B_G to zero. Classical accuracy C_A, by contrast, approaches C_{sp} as $P \to 0$ and C_{se} as $P \to 1$, remaining bounded away from zero and thus failing to signal the collapse of predictive utility at prevalence extremes.

Property 2 —Symmetry under class exchange

Under the transformation $(P, C_{se}, C_{sp}) \to (1-P, C_{sp}, C_{se})$, the roles of *PPV* and *NPV* exchange, but their product is invariant. Therefore, $B_G(P, C_{se}, C_{sp}) = B_G(1-P, C_{sp}, C_{se})$; thus, the index is symmetric under the simultaneous relabeling of classes and complementing of prevalence. This means B_G treats both classes on equal footing regardless of which is designated as positive.

Property 3 — Conservative bound through arithmetic and geometric mean (AM-GM)

By applying the arithmetic–geometric mean inequality applied to *PPV* and *NPV*,

$$B_G = \sqrt{B_{se}B_{sp}} \leq \frac{B_{se}+B_{sp}}{2} = \frac{B_J+1}{2}, \tag{12}$$

with equality if and only if $B_{se} = B_{sp}$. Thus B_G lies strictly below the arithmetic mean of the predictive values whenever they are asymmetric, penalizing classifiers whose positive and negative predictions have different reliability. This makes B_G a conservative companion to the Bayesian Youden index B_J: it rewards not only high predictive values but also balances between them.

Property 4 —Unique interior maximum

B_G vanishes at both extremes of (0,1) and is strictly positive on the interior, thus, it attains a unique global maximum through continuity. This optimal prevalence represents the operating point at which the classifier's joint predictive performance is maximized, and its derivation is the content of the proposition in Section IV.B. Notably, this maximum does not occur at $P = 0.5$ in general but rather shifts toward the class for which the classifier is more reliable, providing a natural data-driven criterion for assessing the practical operating range of the model.

Property 5 —Nonlinearity and prevalence sensitivity

Unlike C_A, which is linear in P and insensitive to prevalence shifts away from the training distribution, B_G is a rational function of P whose numerator grows as $P(1-P)$ and whose denominator AB depends on P through the classifier's marginal prediction probabilities. This double dependence on prevalence, once through the explicit $P(1-P)$ factor and once through A and B, produces a strictly concave shape of B_G over (0,1) and ensures that any departure from the optimal prevalence in $(0,1)$ is immediately reflected in a reduction of the metric, a penalization mechanism entirely absent from C_A.

III. Algebraic Equivalence between Classical and Bayesian Geometric Accuracy

A. Equivalence Condition

Having established that classical accuracy C_A and Bayesian geometric accuracy B_G respond differently to prevalence, the former remaining linear and insensitive to extreme regimes (Property 5) and the latter vanishing at boundaries and attaining a unique interior maximum (Property 4), a natural question arises: for which values of prevalence P do both measures agree? The answer to this question is not merely of theoretical interest; it delimits the operating interval within which relying on C_A produces an optimistic and misleading assessment of true predictive performance. The equivalence condition is $C_A = B_G$; that is,

$$PC_{se} + (1-P)C_{sp} = \sqrt{\frac{P(1-P)C_{se}C_{sp}}{AB}}. \tag{13}$$

Squaring both sides, which is valid because both sides are nonnegative on (0,1), and multiplying through by AB yield

$$\left(PC_{se} + (1-P)C_{sp}\right)^2 AB = P(1-P)C_{se}C_{sp}. \tag{14}$$

Expanding the left-hand side using the expressions for A and B and collecting all terms on one side produces a polynomial equation in P. Noting that A and B are each linear in P, the product AB is quadratic in P, and $\left(PC_{se} + (1-P)C_{sp}\right)^2$ is also quadratic in P, the full left-hand side appears to be degree 4 in P. The right-hand side is degree 2; however, upon full expansion the coefficient of P^4 cancels exactly, and their difference defines a degree 3 polynomial:

$$f(P) = \left(PC_{se} + (1-P)C_{sp}\right)^2 AB - P(1-P)C_{se}C_{sp}, \tag{15}$$

whose coefficients, expressed in terms of $d = C_{se} - C_{sp}$ and C_J, are

$$\beta_3: dC_J(1-C_J);\ \beta_2: C_J\left[d(2C_{sp}-1) + C_{sp}(2-C_J)\right] - C_{sp}(d+1);\ \beta_1: C_{sp}(1-C_{sp})(d+2-2C_J);\ \beta_0: -C_{sp}\left(1-C_{sp}\right)^2.$$

This polynomial has exactly two real roots in (0,1) for any classifier with $C_J > 0$, which are the two critical prevalence values at which classical and Bayesian accuracies coincide. A third real root always exists but lies outside (0,1) and has no practical interpretation for the potato crop classifier.

Using Eq. (15), $f(0) = -C_{sp}\left(1-C_{sp}\right)^2 < 0$ and $f(1) = -C_{se}(1-C_{se})^2 < 0$, both strictly negative for any nondegenerate classifier. Because $f(0) < 0$ and $f(1) < 0$, the polynomial must cross zero an even number of times on (0,1). As a cubic whose leading coefficient β_3 is nonzero,

$f(P)$ must attain positive values somewhere in the interior, guaranteeing exactly two real roots in (0,1) for any classifier with $C_J > 0$ and $C_{sp} \neq C_{se}$.

B. Structure of the Cubic Polynomial and Its Roots

The polynomial $f(P)$ defined in Eq. (15) is of degree 3, with the coefficients given in Section III.A. Unlike a squared equation, the boundary points $P = 0$ and $P = 1$ are not roots of $f(P)$. Direct substitution confirms this: at $P = 0, f(0) = \beta_0 < 0$ for any classifier with $C_{sp} \in (0,1)$, and at $P = 1$, $f(1) = \beta_3 + \beta_2 + \beta_1 + \beta_0 \neq 0$, in general.

The cubic $f(P)$ has exactly three real roots. For any classifier with $C_J > 0$, exactly two of these roots lie in the interior (0, 1) and constitute the two prevalence points of genuine practical interest. The third root is always real but lies outside (0, 1) and has no practical interpretation. The two interior roots are obtained directly from the cubic using standard numerical methods or, in principle, through Cardano's formula applied to the coefficients derived in Section III.A.

C. Two Critical Prevalence Values

The two roots of $f(P)$ in (0, 1) do not admit simple closed-form expressions analogous to the quadratic formula, because $f(P)$ is a cubic polynomial. They are obtained from the coefficients derived in Section III.B by standard numerical root-finding, or, in principle, through Cardano's formula applied to:

$$f(P) = \beta_3 P^3 + \beta_2 P^2 + \beta_1 P + \beta_0 = 0. \quad (16)$$

The existence and uniqueness of exactly two roots in (0, 1) are guaranteed by the boundary result established in Section III.B. The two roots satisfy $P^*_{A-1} < P^*_{A-2}$ and their separation $\Delta = P^*_{A-2} - P^*_{A-1}$ increases with the discriminative capacity of the classifier, as measured by C_J: a wider interval signals a stronger classifier, whereas the interval collapses as $C_J \to 0$.

D. Interpretation of Critical Prevalence

The two prevalence P^*_{A-1} and P^*_{A-2} are the only points in (0, 1) at which classical accuracy and Bayesian Geometric Accuracy agree. Their separation and location have direct diagnostic information about the classifier; however, the direction of the overestimation depends on the relative order of sensitivity and specificity.

For $P \in I = (P^*_{A-1}, P^*_{A-2})$, one can verify that $C_A > B_G$: classical accuracy overestimates the true joint predictive performance captured by B_G. This is precisely the regime in which relying on C_A is most dangerous in practice. It corresponds to prevalence values where the classifier performs better than it does from a predictive standpoint. The interior interval I is where $B_G > C_A$, meaning the Bayesian metric provides a more favorable and honest characterization. This pattern holds for the potato crop classifier: C_A starts near $C_{sp} \approx 1$ at $P = 0$, whereas B_G starts at zero; thus, C_A dominates at low and high prevalence values, and B_G only exceeds C_A within the interior interval. When $C_{se} > C_{sp}$ the roles reverse: C_A starts below B_G for low prevalence, and the overestimation region I becomes the interior interval. In both cases the interval remains the central reference region, but its interpretation as an overestimation or underestimation regime for C_A reverses depending on which intrinsic metric dominates.

The distance between the two critical prevalence values (Δ) is a strictly increasing function of C_J. It vanishes as $C_J \to 0$, when the classifier has no discriminative information and both metrics collapse toward the same trivial behavior, consistent with Property 1, it expands as the classifier improves, reaching its maximum for a perfect classifier. This means that the width of the interval $\mathcal{I}$ is itself a measure of discriminative quality; a wider interval indicates a stronger classifier, whereas a narrow interval indicates that classical and Bayesian accuracy remain close across most of the prevalence range and that the risk of overestimation is limited.

Finally, P^*_{A-1} coincides with the unique maximizer of B_G established in Property 4: it is simultaneously the prevalence at which both metrics first agree and the prevalence at which the Bayesian metric attains its peak predictive performance. P^*_{A-2} marks the upper boundary beyond which C_A again dominates, and its location is governed by the overall discriminative strength of the classifier as captured by C_J. Both roots are obtained from the cubic polynomial $f(P)$ derived in Section III.B and depend on C_{se}, C_{sp} and C_J in a way that reflects the full nonlinear structure of the Bayesian updating. Unlike the simpler closed-form expressions available for related metrics, no reduction to a quadratic is possible here.

IV. Critical Prevalence and Their Connection to the Bayesian Youden Index

A. Operating Interval and the Overestimation Region

The two critical prevalence values $P^*_{A-1}$and P^*_{A-2} derived in Section III partition the unit interval (0, 1) into three regions with qualitatively different behaviors, and the direction of the overestimation depends on whether $C_{se} > C_{sp}$ or $C_{se} < C_{sp}$.

When $C_{se} > C_{sp}$: for $P \in (0, P^*_{A-1})$ and $P \in (P^*_{A-2}, 1)$, B_G exceeds C_A, meaning that the Bayesian metric is more favorable than classical accuracy; for $P \in I$, $C_A > B_G$ and classical accuracy overstates true joint predictive performance. This is the overestimation region for this case. When $C_{se} < C_{sp}$: the roles of the regions invert. C_A starts near C_{sp} at $P = 0$, whereas B_G starts at zero, thus, $C_A > B_G$ for $P \in (0, P^*_{A-1})$ and $P \in (P^*_{A-2}, 1)$. For $P \in I$, $C_A < B_G$ and the Bayesian metric provides the most accurate assessment. Therefore, the overestimation region is $(0, P^*_{A-1}) \cup (P^*_{A-2}, 1)$, covering both tails of the prevalence range, and the safe operating interval where $B_G \geq C_A$ is the interior $\mathcal{I}$.

In both cases, a classifier operating at a prevalence outside the interval $\mathcal{I}$ risks having its performance misrepresented by C_A, and the width □ remains the principled measure of discriminative strength established in Section III.D.

The width of this region, □, is not arbitrary. As established in Section III.B, it is a strictly increasing function of C_J. Therefore, a wider overestimation interval therefore signals a more discriminative classifier, which may seem counterintuitive: stronger classifiers produce larger regions where C_A overestimates B_G. This is because a stronger classifier has more asymmetric predictive values across the

prevalence range; thus, the gap between the linear behavior of C_A and the concave behavior of B_G (Property 5) is more pronounced. For a near-trivial classifier with $C_J \to 0$, both metrics converge and the interval collapses, consistent with Property 1.

B. The Optimal Prevalence P^*_{A-1} and the Optimal Prevalence Proposition

By Property 4, B_G attains a unique global maximum at some interior prevalence $P_{opt} \in (0,1)$. To find this point, we differentiate B_G with respect to P and set the derivative to zero. Writing

$$B_G^2 = h(P) = P(1-P)\frac{C_{se}\,C_{sp}}{AB} \tag{17}$$

is equivalent to maximize $h(P)$ because the square root is monotonically increasing. Applying the quotient rule and using $AB = -C_J^2P^2 + C_J\big(2C_{sp} - 1\big)P + C_{sp}(1 - C_{sp})$, the first-order condition becomes

$$dC_JP^2 - 2uP + u = 0 \tag{18}$$

with $d = C_{se} - C_{sp}$ and $u = C_{sp}\big(1 - C_{sp}\big)$.

Proposition. *For a classifier with C_{se}, $C_{sp} \in$ (0,1) and $C_J > 0$, the unique* prevalence *maximizing B_G is*

$$P_{opt} = \frac{u - \sqrt{u(u - dC_J)}}{dC_J}. \tag{19}$$

Proof. Eq. (19) is obtained by applying the quadratic formula to Eq. (18) and selecting the root in (0, 1). The second root is outside (0, 1) and is discarded.

A remarkable observation follows immediately: P_{opt} coincides exactly with the critical prevalence P^*_{A-1}, the smaller root of the cubic $f(P)$ derived in Section III.B. This is not a coincidence because P^*_{A-1} is simultaneously the prevalence at which $C_A = B_G$ on the boundary of the overestimation region and the prevalence at which B_G achieves its global maximum. The classifier's joint predictive performance peaks precisely at the point where classical accuracy transitions from overestimating to correctly representing it.

C. Connection to the Bayesian Youden index

The Bayesian Youden index is defined as $B_J = B_{se} + B_{sp} - 1$. In general, these equilibrium prevalence values do not coincide with the critical prevalence P^*_{A-1} and P^*_{A-2} of B_G. This is not surprising because $B_J = PPV + NPV$ and $B_G = \sqrt{PPV.NPV}$ aggregate the same predictive values through fundamentally different operations, additive versus geometric, and their respective equilibrium structures with classical metrics reflect these different sensitivities to the asymmetry between *PPV* and *NPV*.

Both metrics share is a qualitative dependence on prevalence that classical accuracy entirely lacks. B_J and B_G are nonlinear functions of P, both vanish at the boundary extremes $P = 0$ and $P = 1$ consistent with Property 1, and both identify prevalence regimes where classical accuracy produces a misleading assessment of true predictive performance. The equilibrium prevalence of B_G, the roots P^*_{A-1} and P^*_{A-2} of the cubic $f(P)$ derived in Section III.B and the equilibrium prevalence of B_J together provide a richer characterization of the prevalence range than either alone. A practitioner equipped with both sets of equilibria can identify the full spectrum of operating conditions under which Bayesian metrics diverge meaningfully from classical accuracy and can select the metric whose structural properties best align with the deployment context.

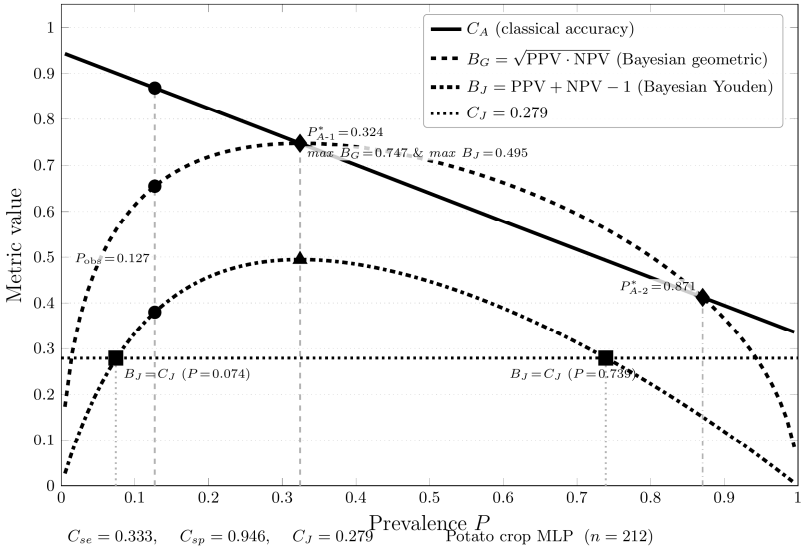

Fig. 1. C_A, B_G, B_J and C_J as functions of prevalence: shared and exclusive equilibria.

The deeper connection between B_G and B_J lies in the inequality established in Property 3, with equality if and only if $PPV = NPV$. This means B_G is always a conservative lower bound on the arithmetic mean of the predictive values, and the gap between them reflects the degree of asymmetry between the positive and negative predictive reliability of the classifier across the prevalence range.

Fig. 1 shows the classical accuracy C_A, B_G, B_J, and classical Youden index $C_J = 0.279$ (horizontal dashed-dotted line) as functions of prevalence P, for the potato crop MLP classifier. Diamond markers indicate the equilibria of $C_A = B_G$ at $P^*_{A-1} = 0.324$ and $P^*_{A-2} = 0.871$, derived as the roots of the cubic polynomial in Section III.B. Square markers indicate the equilibria of $B_J = C_J$ at $P = 0.074$ and $P = 0.739$. The four

equilibrium points are all distinct, confirming that B_G and B_J define complementary and nonoverlapping partitions of the prevalence range. P^*_{A-1} and P^*_{A-2} simultaneously mark the maximum of $B_G = 0.747$ and the maximum of $B_J = 0.495$, identified in the proposition in Section IV.B. The observed prevalence $P = 0.127$ (teal dashed line) is shown with its corresponding values on all three metric curves. The data associated with the presented values are explained in Section V.

V. Motivating Application: Crop Disease Detection from Multispectral Images

A. Study Context

The motivating application is the discrimination of healthy and diseased potato crop plants from multispectral indices. The identification of diseases and physiological disorders in potato crops from unmanned aerial vehicles is an active field of research. A recent study [16], classified six severity levels of *Verticillium* wilt in potatoes using a dataset with n = 212 samples at 1 m×1 m sites. A six-level damage severity scale was used for each sample considering the degree of leaf damage and the stratum in which the symptoms occurred. The six levels correspond to *0*, Healthy plant; (1). plant affected between 0% and 25% with chlorotic leaves in transition to necrosis in the lower part; (2). plant affected between 25% and 50%, with chlorotic and necrotic leaves in the middle part of the plant and wilting in the lower part; (3). plant affected between 50% and 90% with chlorotic and necrotic leaves in the upper part of the plant, wilting in the middle and lower parts of the secondary stems; (4). plant affected at 90% with several leaves at its terminal end showing a chlorotic and necrotic appearance, as well as chlorotic and wilted stems with little stability; and (5). dead plant. In this study, we converted the original dataset into a binary partition (*0 = healthy*, *1 = diseased*) with a high imbalance, yielding 185 and 27 samples for absence and presence classes, respectively, corresponding to an observed prevalence of $P = \frac{27}{212} \approx (0.127)$. Predictors consist of four multispectral indices: normalized difference vegetation index (*NDVI*), enhanced vegetation index (*EVI)*, normalized difference red edge (*NDRE)*, and green leaf index (*GLI*), as well as *plant height* derived from the sensor on board of the aerial platform.

B. Model and Confusion Matrix

Table 2. Confusion matrix for the multilayer perceptron classifier trained on multispectral indices

	Real absence	Real presence
Predicted absence	175	18
Predicted presence	10	9

An MLP was trained using standard backpropagation with a two-hidden layer architecture of sizes (10, 5), learning rate of 0.05, and 500 maximum iterations. Table 2 shows the resulting confusion matrix.

The intrinsic classifier parameters from the confusion matrix are sensitivity $C_{se} = \frac{9}{27}(0.333)$, specificity $C_{sp} = \frac{175}{185}(0.946)$, and the Youden index $C_J = 0.279$. The prevalence observed in the sample is $P = 0.127$. This classifier falls in the case $C_{se} < C_{sp}$, thus, the overestimation region identified in Section IV.A covers the tails $(0, P^*_{A-1}) \cup (P^*_{A-1}, 1)$, and the safe operating interval where $B_G \leq C_A$ is the interior $\mathcal{J}$.

A Shiny application developed in R is available on the GitHub repository of the corresponding author (https://github.com/darghan/unal/blob/main/accuracyGeom.R) to facilitate reproducibility and interactive exploration of the framework. The application allows the reader to vary sensitivity, specificity, and observed prevalence through sliders and observe in real time the behaviors of C_A and B_G as functions of prevalence, the location of the two critical prevalence values, and the optimal operating point. The repository also includes figures presented in this document and a comparative numerical table of B_G against the scalar alternatives F1-score, Matthews correlation coefficient (MCC), and G-mean evaluated at the confusion matrix of the potato crop classifier, along with a summary of their formal structural properties.

C. Critical Prevalence and Position of the Observed Prevalence

Applying the cubic polynomial $f(P)$ derived in Section III.B to the classifier parameters, the $P^*_{A-1} = 0.324$ and $P^*_{A-2} = 0.871$are the two critical prevalence values. The observed prevalence $P = 0.127$ falls strictly below P^*_{A-1}, placing it in the overestimation region identified in Section IV.A where $C_A > B_G$. At this prevalence the two metrics are $C_A = 0.868$ and $B_G = 0.655$, a gap of 0.213. The classical accuracy of 0.868 may give an incorrect sense of high precision and lead the researcher to trust the proposed model; however, B_G is undoubtedly much more realistic because of the predictive nature of the proposal, because it is seriously penalized by the class imbalance. The classifier correctly identifies only 9 of 27 disease-present cases, and at a prevalence of 12.7% the PPV is severely diminished by the rarity of the positive class, a failure that C_A conceals and B_G exposes.

The width of the overestimation interval $P^*_{A-2} - P^*_{A-1} = 0.547$ has an additional meaning established in Section IV.B: it is the unique prevalence at which B_G attains its global maximum (0,1). Evaluating B_G at this optimal prevalence $B_G(P^*_{A-1}) = 0.747$. The classifier's joint predictive performance would peak at $B_G = 0.747$ if deployed in a population with a disease prevalence of 32.4%. However, the observed prevalence of 12.7%, is less than half that of the optimal value, indicating that the model operates in a suboptimal prevalence regime where its predictive capacity is substantially below its theoretical maximum. The gap between $B_G(P^*_{A-1})$ and $B_G(0.127) = 0.655$ quantifies this lost predictive potential. A deployment context with a prevalence closer to 0.324—for instance, during periods of active disease outbreak—would yield $B_G = 0.747$ under the fixed-classifier assumption, recovering 0.092 units of joint predictive performance relative to the observed operating point. This comparison is conditional on the classifier's intrinsic parameters remaining fixed, which is a theoretical benchmark rather than an empirical prediction: in practice, a classifier retrained or recalibrated at a different prevalence would produce different values of C_{se} and C_{sp}, and its B_G curve would differ accordingly. The value of this benchmark lies precisely in isolating the prevalence effect from the retraining effect, allowing both to be independently assessed.

Fig. 2 shows the classical accuracy C_A and Bayesian geometric accuracy B_G as a function of prevalence P, for the

potato crop MLP with sensitivity $C_{se} = 0.333$, $C_{sp} = 0.946$, and the Youden index $C_J = 0.279$, $n = 212$. Because $C_{se} < C_{sp}$, the overestimation region where $C_A > B_G$ covers both tails of the prevalence range (shaded red), whereas $B_G > C_A$ only in the interior interval $\mathcal{I}$ (shaded green). The two equilibrium prevalence $P^*_{A-1} = 0.324$ and $P^*_{A-2} = 0.871$ are the roots of the cubic polynomial derived in Section III.B where both metrics coincide. P^*_{A-1} simultaneously serves as the optimal prevalence P_{opt} at which B_G achieves its global maximum of 0.747 (proposition in section 4.2). The observed prevalence $P = 0.127$ falls in the left overestimation tail, where $C_A = 0.868$ and $B_G = 0.655$, a gap of 0.213 that quantifies the overestimation of predictive performance produced by class imbalance. The horizontal dotted lines indicate the intrinsic classifier parameters.

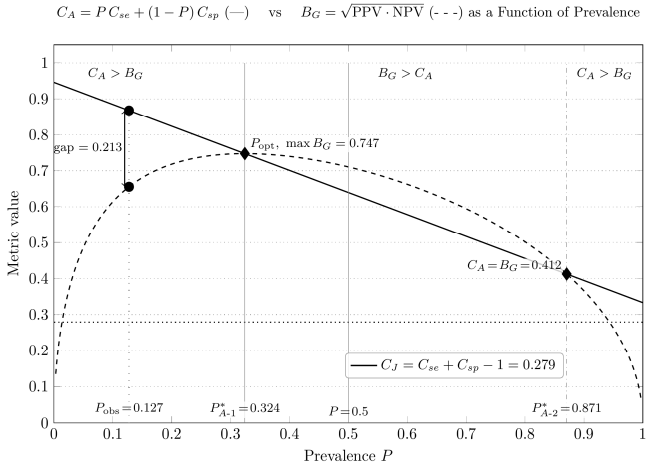

Fig. 2. Classical accuracy C_A and Bayesian geometric accuracy B_G as a function of prevalence P for the potato crop MLP.

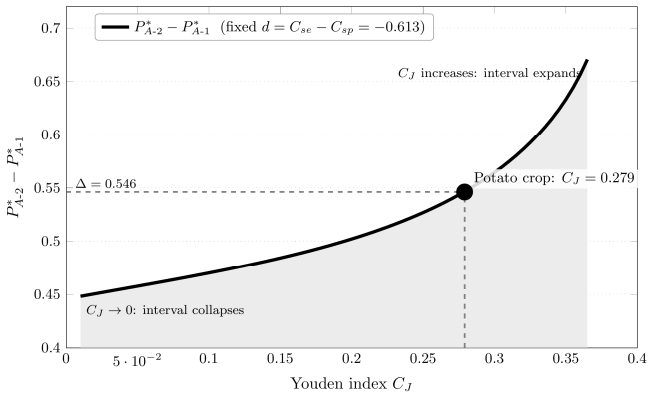

Fig. 3. Width of the divergence interval $\mathcal{I}$ as a strictly increasing function of the Youden index C_J computed with fixed asymmetry.

Fig. 3 shows the width of the divergence interval $\mathcal{I}$ as a strictly increasing function of the Youden index C_J, computed with fixed asymmetry $d = C_{se} - C_{sp} = -0.613$ (the same asymmetry as the potato crop classifier). The width vanishes as $C_J \to 0$, consistent with Property 1, and grows monotonically as discriminative capacity improves, confirming that the separation between the two critical prevalence values constitutes a principled measure of classifier quality. The highlighted point corresponds to the potato crop MLP ($C_J = 0.279$, Δ=0.546), a moderate discriminative strength classifier whose divergence interval covers a substantial portion of the prevalence range relevant to agricultural disease monitoring.

D. Bootstrap Assessment of the $C_A - B_G$ gap

To assess whether the divergence $C_A - B_G = 0.213$ observed at the sample prevalence of 12.7% is a stable feature of the classifier or an artifact of the specific data partition, a nonparametric case-level bootstrap was performed. The 212 observations were reconstructed from the confusion matrix entries: 9 TP, 10 FP, 18 FN and 175 TN — and resampled with replacement B = 5000 replicates. For each bootstrap replicate, the confusion matrix was recomputed from the resampled cases, new values of sensitivity and specificity were derived, and C_A and B_G were evaluated at the observed prevalence $P = 0.127$. No distributional assumption was imposed on any metric: the empirical resampling of the case pool entirely explains the bootstrap distributions. Table 3 presents the bootstrap mean, standard deviation and 95% percentile confidence interval for C_A and B_G and their differences.

Table 3. Bootstrap confidence intervals for C_A, B_G and the gap C_A-B_G at $P = 0.127$ (*B B = replicates, seed = 2024*)

Metric	Observed value	Bootstrap mean	SD	95% CI
C_A	0.868	0.868	0.018	[0.831,0.903]
B_G	0.655	0.652	0.079	[0.488,0.797]
$C_A - B_G$	0.213	0.216	0.061	[0.106,0.347]

VI. Discussion

Before discussing the implications of the framework, it is useful to situate B_G relative to the most widely used alternatives for imbalanced classification: the F1-score, the MCC, and the G-mean. All three are scalar summaries computed from confusion matrix counts at a fixed operating point and a fixed class distribution. They do not express prevalence as a formal argument; therefore, they cannot answer the question of how performance changes as the deployment prevalence shifts away from the training distribution. The G-mean, which is the geometric mean of sensitivity and specificity, is the closest structural analog to B_G. It operates in the intrinsic space of C_{se} and C_{sp} — answering whether the classifier discriminates, rather than in the predictive space of *PPV* and *NPV*, answering whether predictions can be trusted. This distinction is not merely semantic: a classifier with high G-mean but low prevalence will have a severely diminished *PPV* regardless of its discriminative capacity, a failure that G-mean cannot detect and B_G exposes by construction. MCC incorporates all four cells of the confusion matrix and correlates with BG at the observed prevalence, but it provides no mechanism for identifying the prevalence regimes where the evaluation is reliable or misleading. B_G complements rather than replaces these metrics. It adds a prevalence-sensitive, prediction-space perspective that scalar summaries cannot structurally provide and uniquely generates an algebraically grounded operating interval that identifies which prevalence regimes render classical accuracy misleading.

The B_G metric complements rather than replaces these metrics: it adds a prevalence-sensitive, prediction-space perspective that scalar summaries structurally cannot provide, and uniquely generates an algebraically grounded operating interval that identifies which prevalence regimes render

classical accuracy misleading. Table 4 summarizes the key formal properties that distinguish B_G from the three main alternatives.

Table 4. Formal properties of B_G against scalar alternatives for imbalanced binary classification

Property	F1	MCC	G-mean	B_G
Prevalence as a formal argument	No	No	No	Yes
Vanishes at $P \to 0$ and $P \to 1$	No	No	Partially	Yes
Anchored in the prediction space (*PPV*/*NPV*)	No	No	No	Yes
Symmetric under class exchange	No	No	Yes	Yes
Generates algebraic operating interval	No	No	No	Yes
Closed-form prevalence curve	No	No	No	Yes

No scalar alternative among the F1-score, MCC, and G-mean simultaneously satisfies all six properties. B_G is the only metric in this comparison that expresses performance as a function of prevalence rather than a point estimate, making it an appropriate tool when the deployment distribution is expected to differ from the training distribution. The structural affinity between B_G and the G-mean is not coincidental. Both are geometric means, and this shared construction endows them with two properties that arithmetic-based metrics, such as the F1-score and MCC, do not possess. First, the geometric mean vanishes when either of its factors approaches zero, correctly signaling a total failure of one component even when the other is high —a behavior that arithmetic averaging masks by compensating low values with high ones. Second, the geometric mean more aggressively penalizes asymmetry between its factors than any arithmetic mean, rewarding classifiers that consistently perform across both classes rather than excelling on one at the expense of the other. Where B_G and the G-mean diverge is in what they average: G-mean takes the geometric mean of C_{se} and C_{sp}, summarizing discriminative balance in the intrinsic space, whereas B_G takes the geometric mean of PPV and NPV, capturing predictive balance in the deployment space—the same geometric principle applied one Bayesian inversion deeper.

Data analysis in supervised machine learning contexts with imbalanced classes is a major challenge, even when using conventional algorithmic performance metrics. Class imbalance often limits the practical usefulness of classification models by biasing classifiers toward the majority class — a limitation that persists in both conventional machine learning and deep learning algorithms [17]. The intrinsic tendency of traditional learning algorithms to favor the most common classes and mask the contribution and potential information value of the underrepresented class poses a challenge [18]. Imbalanced learning remains an active field of research, encompassing topics ranging from data preprocessing and classifier modification to algorithmic parameter optimization. Therefore, developing metrics and learning paradigms capable of explicitly accounting for distributional asymmetry is of particular importance, as it determines whether classification predictions remain reliable across different prevalence regimes [18].

The problem of imbalanced data classification is common in real-world applications, such as detecting anomalies in binary time series [19], diagnosing crop disease [20, 21], and technology adoption studies [22]. In these fields, such an imbalance emerges naturally when data collection occurs in the early stages of an event of interest, a phase typically characterized by a predominance of the negative class over the positive instances. For instance, early monitoring of crop disease progression captures predominantly healthy samples with absence outnumbering presence [21, 23]. Classifiers overfit to the majority class in modeling, leading to unstable and even false performance in the algorithm performance evaluation.

The selection of evaluation metrics in the presence of imbalanced classes does not have a universal answer: the most appropriate metric depends on the specific objective of the problem and the user's priorities. This conclusion, derived from a comparative analysis on multiclass classification with imbalanced data, underscores the need to continue exploring and testing algorithmic performance metrics beyond classic metrics such as accuracy or F1-score, which may be insufficient or misleading depending on the context. For example, recent studies indicate that metrics such as the Matthews Correlation Coefficient (MCC) and G-mean are often more informative alternatives under imbalanced class distributions [20]. In class-imbalance scenarios—common in crop disease detection, medical diagnosis, and system monitoring—the wrong choice of metric can lead to incorrect conclusions about the actual capacity of the classifier, which justifies the development and evaluation of alternative metrics that more accurately capture performance in the minority class [24].

The results of this study indicate that the class imbalance in binary classification is not a uniformly pathologic condition, but rather a prevalence-dependent phenomenon with identifiable tolerance thresholds. The two critical prevalence values derived from the equivalence condition between classical and Bayesian geometric accuracy delimit a region within which the Bayesian metric exceeds classical accuracy, indicating that not all degrees of imbalance are equally problematic. For the potato crop classifier, this safe operating interval spans from approximately 32% to 87% disease prevalence, a range that, altough not observed in the current sample, is not unrealistic during active outbreak periods. This reframes the class imbalance problem: rather than declaring any deviation from balance as unacceptable, the framework provides an algebraically grounded criterion for determining which prevalence regimes are tolerable from a predictive standpoint and which are not.

The decision to anchor this criterion on C_A and B_G rather than on classical and Bayesian Youden indices is deliberate. Both pairs of metrics generate their own equilibrium structures, and a practitioner seeking a single operating threshold faces a choice between them. The accuracy-based equilibria derived here reflects the question that matters most at deployment: given a prediction, how often is it correct? This is a more operationally relevant question than the balanced sensitivity–specificity tradeoff captured by Youden-type indices, which evaluate the classifier from the perspective of someone who already knows the true label. However, the framework leaves open a natural extension: rather than selecting one pair of equilibria, a practitioner could define an operating region that simultaneously

accounts for multiple equilibrium points from both predictive and discriminative perspectives. A principled selection criterion might favor prevalence values closest to 0.5, where both classical and Bayesian metrics converge and the classifier operates under the most balanced conditions. This approach replaces the arbitrary 50-50 split convention with a data-driven tolerance region derived from intrinsic parameters of the classifier.

The divergence $C_A - B_G$ itself deserves recognition as a complementary diagnostic measure. Rather than selecting one metric over the other, reporting the gap between them as a function of the observed prevalence provides a single interpretable quantity that summarizes how much the classifier's predictive reliability deviates from what classical accuracy suggests. A large gap indicates that the class imbalance is actively distorting the evaluation; a small gap indicates that both perspectives are consistent and that the classical metric can be trusted. This divergence measure can be generalized to any binary classification algorithm, including gradient boosting, support vector machines, random forests, and deep neural networks, as it only depends on the confusion matrix entries and the observed prevalence, which are the quantities available for any classifier regardless of its internal architecture.

The potato crop application illustrates precisely what is at stake when this diagnostic is ignored. The MLP achieved a classical accuracy of 0.868, a value that, taken in isolation, would be conventionally interpreted as evidence of strong model performance and could confidently guide subsequent decisions about which spectral indices explain the presence or absence of the disease. However, a Bayesian geometric accuracy of 0.655 tells a fundamentally different story: a gap of 0.213 reveals that the high classical accuracy is largely an artifact of the severe class imbalance at 12.7% prevalence, where the model correctly classifies the abundant healthy plants while failing on more than two-thirds of the diseased cases it was designed to detect. Therefore, a perception of high accuracy generated by the perceptron under these conditions is misleading, and any variable selection process based on this model's feature importance or prediction outputs of this model would inherit that misleading signal.

The analysis further indicates that a training partition with a prevalence closer to the optimal value of 32.4%, achievable through minority class oversampling, majority class undersampling, or synthetic data augmentation, would place the classifier in a regime where its predictive capacity is substantially closer to its theoretical maximum and where C_A and B_G are in closer agreement. Operating at the observed prevalence of 12.7% means that the model is functioning nearly 20 percentage points below its optimal predictive performance without any change in its parameters, purely because of the prevalence regime. This indicates that rebalancing the training data before fitting the classifier is not merely a computational convenience but a substantive requirement for generating predictions whose reliability can be assessed and trusted for applications involving severely imbalanced classes of this magnitude. Data rebalancing approaches represent a versatile data-level strategy for addressing class imbalance by amplifying the relative representation of the minority class independently of the learning algorithm or classification context [25]. Without such adjustment, the vegetation indices selected as explanatory variables for disease presence or absence would reflect the classifier's ability to identify healthy plants rather than its ability to detect the pathology, leading to potentially incorrect conclusions about the usefulness of multispectral indices for early detection of *Verticillium* wilt in potato crops and undermining the practical value of the remote sensing pipeline that generated the data. Note that this recommendation operates on a different level from the theoretical benchmark established in Section 5.3, where the improvement in B_G was computed under fixed intrinsic parameters, isolating the pure effect of prevalence on the metric. In this study, the suggestion is to retrain the classifier under a more balanced distribution, which will generally alter C_{se} and C_{sp} themselves and produce a different B_G curve entirely. Both analyses are complementary: the fixed-classifier benchmark quantifies what is lost by operating at a suboptimal prevalence regime, whereas the retraining recommendation addresses the separate question of whether the model's intrinsic discriminative capacity can be improved. The distinction matters because a rebalanced classifier with higher sensitivity but lower specificity may or may not improve B_G at the target prevalence, depending on which intrinsic parameter changes more —a question that requires empirical validation and lies beyond the scope of the present theoretical framework.

VII. Conclusions

Bayesian geometric accuracy provides a principled and operationally meaningful alternative to classical accuracy for evaluating binary classifiers under class imbalance. Unlike classical accuracy, which is linear in prevalence and insensitive to the rarity of the positive class, the new measure is nonlinear, vanishes at both prevalence extremes, and penalizes the inflated performance figures that classical accuracy produces when the majority class dominates by construction. The equivalence condition yields roots that delimit the operating interval within which classical accuracy overestimates true predictive performance, and their separation constitutes a quantitative measure of discriminative strength that requires nothing beyond the entries of the confusion matrix. Together, these results replace the heuristic practice of flagging any class imbalance as problematic with an algebraically grounded criterion that precisely identifies which prevalence regimes are tolerable from a predictive standpoint and which are not.

Applied to an MLP trained on spectral indices for the binary classification of *Verticillium* wilt presence in potato crops, the framework reveals that the classical accuracy of 0.868 substantially overstates the predictive reliability of the model at the observed prevalence of 12.7%, where $B_G = 0.655$ and the gap of 0.213 is directly attributable to severe class imbalance. The classifier is operating nearly 10 percentage points below its theoretical maximum predictive performance, which would be attained at a prevalence of 32.4%, a value achievable through rebalancing of the training data before model fitting. Without such adjustment, the spectral indices selected as explanatory variables for disease presence or absence reflect the model's ability to classify the abundant healthy class rather than its capacity to detect the pathology, undermining the validity of any subsequent

variable selection or biological interpretation. On the basis of this insight, the Bayesian geometric accuracy defines a safe space of imbalance ratios that can be effectively managed, enabling more reliable model development and guiding the design of innovative data augmentation or resampling strategies tailored to agricultural disease monitoring.

The results establish three connected claims. Bayesian geometric accuracy is a principled prevalence-sensitive metric that operates in the deployment-relevant space of predictive values, vanishes at both prevalence extremes, and penalizes class imbalance by construction. The equivalence condition with classical accuracy yields a cubic polynomial whose two interior roots delimit a safe operating interval: imbalance is tolerable precisely when the observed prevalence falls within that interval, and its width is a monotone increasing function of discriminative strength as measured by the Youden index. The bootstrap assessment confirms that the divergence $C_A - B_G = 0.213$ at the observed prevalence of 12.7% is not an artifact of the specific data partition—across 5000 case-level resamples, the gap remained strictly positive in every replicate, with a 95% confidence interval of [0.106, 0.347]. The wider bootstrap variability of B_G relative to C_A reflects the greater sensitivity of predictive values to minority class fluctuations, confirming that B_G is the more reactive and therefore the more honest diagnostic under severe class imbalance.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

AD: Conceptualization, methodology, formal analysis, writing original draft; IL: investigation, resources, data curation and validation, writing original draft; CR: formal analysis and visualization, writing original draft; LC: writing original draft, data curation and validation, writing-review & editing; JJ: formal analysis, visualization, writing original draft. All authors had approved the final version.

References

[1] O. Colliot, *Machine* Learning *for Brain Disorders*. Humana Press, 2023.

[2] M. Altalhan, A. Algarni, and M. T. Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.0429000

[3] M. Kuhn, D. Vaughan, and E. Hvitfeldt, "yardstick: Tidy characterizations of model performance," 2025.

[4] M. Kuhn. (2024). caret: Classification and regression training. [Online]. Available: https://cran.r-project.org/web/packages/caret/index.html

[5] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] K. Beckh *et al.*, "Explainable machine learning with prior knowledge: An overview," ResearchGate, 2021.

[8] E. Lesaffre and A. B. Lawson, *Bayesian Biostatistics*. Wiley, 2012.

[9] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[10] K. M. Sujon, R. Hassan, K. Choi *et al.*, "Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J Big Data*, vol. 12, p. 268, 2025.

[11] S. Sathyanarayanan and B. R. Tantri, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, vol. 27, pp. 4023–4031, 2024. doi: 10.53555/AJBR.v27i4S.4345

[12] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Comput Stat*, vol. 39, pp. 363–383, 2024. doi: 10.1007/s00180-022-01301-9

[13] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: Simple definition and calculation of accuracy, sensitivity and specificity," *Emergency (Tehran)*, vol. 3, pp. 48–49, 2015.

[14] S. Stevič, "Geometric mean," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer, 2025. doi: 10.1007/978-3-662-69359-9_256

[15] L. da Fontoura Costa. (2023). By means of the means: Arithmetic, harmonic, geometric. [Online]. Available: https://hal.science/hal-04152919/document

[16] I. Lizarazo, J. L. Rodriguez, O. Cristancho, F. Olaya, *et al.*, "Identification of symptoms related to potato Verticillium wilt from UAV-based multispectral imagery using an ensemble of gradient boosting machines," *Smart Agricultural Technology*, vol. 3, p. 100138, 2023. doi: 10.1016/j.atech.2022.100138

[17] K. Ghosh *et al.*, "The class imbalance problem in deep learning," *Machine Learning*, vol. 113, no. 7, pp. 4845–4901, 2022. doi: 10.1007/s10994-022-06268-8

[18] W. Chen *et al.*, "A survey on imbalanced learning: Latest research, applications and future directions," *Artif Intell Rev*, vol. 57, p. 137, 2024. doi: 10.1007/s10462-024-10759-6

[19] G. Princz, M. Shaloo, and S. Erol, "Anomaly detection in binary time series data: An unsupervised machine learning approach for condition monitoring," *Procedia Comput Sci*, vol. 232, pp. 1065–1078, 2024.

[20] A. Jensen, P. Brown, K. Groves, and A. Morshed, "Next generation crop protection: A systematic review of trends in modelling approaches for disease prediction," *Comput Electron Agric*, vol. 234, 2025. doi: 10.1016/j.compag.2025.110245

[21] T. Miftahushudur, H. M. Sahin, B. Grieve, and H. Yin, "Remote sensing," *Remote Sens (Basel)*, vol. 17, 2025. doi: 10.3390/rs17030454

[22] S. Banerjee and S. W. Martin, "A binary logit analysis of factors impacting adoption of genetically modified cotton," *AgBioForum*, vol. 12, no. 2, pp. 218–225, 2008. doi: 10.22004/AG.ECON.37140

[23] C. Hou *et al.*, "Recognition of early blight and late blight diseases on potato leaves based on graph cut segmentation," *J Agric Food Res*, vol. 5, p. 100154, 2021. doi: 10.1016/j.jafr.2021.100154

[24] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, 2023. doi: 10.14569/IJACSA.2023.01406116

[25] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: A review from a data perspective," *Journal of Big Data*, vol. 12, p. 71, 2025. doi: 10.1186/s40537-025-01119-4