Classifying Digital Readiness in Traditional MSMEs: Random Forest Baseline Evaluation and Feature Stability on Imbalanced Data

Amal Khairan[1*], Achmad Fuad[2], Yasir Muin[3], Seh Turuy[4]

[1,2,3,4]Informatics, Faculty of Engineering, Khairun University, Ternate, North Maluku, Indonesia
E-mail: [1*]amalkhairan@unkhair.ac.id, [2]foefoe09@gmail.com, [3]yasirmuin@unkhair.ac.id, [4]seh.turuy@unkhair.ac.id

(Received: 11 Sep 2026, revised: 1 Nov 2026, accepted: 9 Nov 2026)

Abstract

Digital transformation can improve the competitiveness of micro, small, and medium enterprises (MSMEs), but evidence on the operational digital readiness of traditional-market businesses in island-city settings remains limited. This study assessed readiness-class distribution, tested a fixed, untuned Random Forest as a diagnostic baseline against a majority-class comparator, and identified stable non-label predictors. A cross-sectional survey involved 159 traders from three traditional markets in Ternate City. Five ordinal items were used to construct the Digital Readiness Score and were removed from the predictor space to prevent direct target leakage. Model performance was estimated through repeated stratified five-fold cross-validation with ten repetitions. Feature relevance was examined using Gini importance, permutation importance, and ranking stability summarized by a Composite Predictive Ranking Score (CPRS). The classes were strongly imbalanced: 67.92% Low, 26.42% Medium, and 5.66% High. Random Forest achieved an accuracy of 0.7384, 0.4555 balanced accuracy, and 0.4470 macro-F1, exceeding the majority baseline, although it produced no true positives for the High class. Online marketing knowledge, perceived importance and benefits of technology, and online selling knowledge formed the most robust indicator set, remaining the top four under out-of-fold permutation importance and all 66 tested CPRS weightings; smartphone adequacy ranked fifth but was weight-sensitive. The results provide a transparent benchmark for subsequent model optimization and targeted digital-capability interventions, but the baseline is not appropriate for individual readiness decisions.

Keywords: Class Imbalance, Digital Readiness, Feature Importance, MSMEs, Random Forest.

I. INTRODUCTION

Digital transformation increasingly affects how MSMEs communicate with customers, promote products, process transactions, obtain market information, and organize daily decisions. These benefits, however, depend on more than the availability of digital tools. Indonesian and developing-country evidence shows that limited infrastructure, weak digital competence, resource constraints, and insufficient organizational support still restrict MSME digital transformation [1], [2]. Digital readiness is therefore treated here as the initial capacity that enables business owners to understand, accept, and use technology in ways that are relevant to their operations.

Conceptually, digital readiness differs from adoption, maturity, and transformation. Readiness concerns preparedness to initiate and benefit from digital change, whereas maturity reflects the extent to which digital implementation has already progressed [3]. For MSMEs, readiness involves technology, people, processes, strategy, integration, technological sensemaking, agility, implementation, security, culture, governance, cost, and risk considerations [4]-[6]. This multidimensional view is important for traditional-market businesses because ownership of a smartphone or Internet access alone does not indicate whether technology can be applied consistently and safely in business activities.

Traditional-market MSMEs also operate under conditions that can make digitalization uneven. Financial access, performance and effort expectations, and social influence have been associated with digital adoption and business performance among traditional-market enterprises [7]. E-commerce, social media, and digital payments are linked to higher business turnover, while digital bookkeeping remains less common [8]. QRIS adoption is likewise shaped by knowledge, perceived usefulness, ease of use, and risk [9], and digital technologies can alter supplier relationships and transaction efficiency [10]. These findings suggest that readiness assessment should combine knowledge, perceptions,

DOI: 10.34148/teknika.v15i3.1582
TEKNIKA, Volume 15(3), November 2026, pp. 573-582
ISSN 2549-8037, EISSN 2549-8045

CC BY SA

transaction practices, and the capacity to integrate technology into routine business processes.

The local setting adds another layer to this assessment. Ternate is treated in this study as an under-studied island-city context rather than as a causal determinant of readiness. Research on smart-island readiness emphasizes infrastructure, human resources, policy, institutional capacity, and Internet penetration as contextually important conditions [11]. At the enterprise level, the Technology-Organization-Environment perspective similarly frames digital use as an interaction among technological conditions, organizational capability, and environmental support [12]. Evidence from emerging economies also points to internal and external determinants such as devices, skills, innovation, internal rules, and financial resources [13], [14]. Multi-attribute approaches are therefore appropriate for organizing readiness indicators, provided that the initial characteristics and constraints of MSMEs are considered [15], [16].

Based on this literature, the instrument is organized into six literature-derived domains: digital infrastructure and access, technology-use competence, technology implementation in business activities, digital management and understanding, support and environment, and digital trust and risk perceptions. These domains are used to structure the indicators; they are not claimed as empirically validated latent constructs. The class-defining items are also separated from the non-label predictors so that the classifier cannot simply reproduce the rule used to create the outcome. This separation is central to the study's leakage-control strategy.

Most digital-readiness studies use indices, maturity models, technology-acceptance approaches, regression, or structural equation modeling. Such methods are useful for describing conditions and relationships among constructs, but they do not directly test whether a set of indicators can classify unseen observations. Machine learning offers a predictive framework that can learn nonlinear patterns without fully specifying relationships a priori [17], and its use for assessing MSME digital transformation in Indonesia is emerging [18]. The present study therefore extends readiness assessment toward predictive classification while maintaining a clear distinction between readiness and successful digital transformation.

Random Forest was selected as a fixed diagnostic baseline because it handles tabular data, captures nonlinear relationships and interactions, and yields feature-importance estimates [19]. The model is not presented as an optimal algorithm. Its role is to test whether non-label indicators contain predictive information beyond the majority-class rule. This choice is particularly relevant because class imbalance can bias predictions toward the dominant class and distort variable rankings [20]. Alternative strategies such as density-based Random Forest variants and resampling methods can improve minority-class recognition [21], [22], but they were intentionally excluded at the baseline stage so that weaknesses caused by the original class distribution could be observed directly. Variable-importance results were also interpreted cautiously because imbalance can affect feature rankings [23].

A second concern is whether the predictors identified as important remain informative across different validation splits. Feature selection can improve classification efficiency by removing weak predictors [24], while recent studies increasingly combine predictive modeling with explainable methods such as SHAP [25]. Readiness-oriented applications have also linked machine-learning predictions with organizational change factors and intervention requirements [26]. In this study, however, the objective is more limited: to identify globally informative and relatively stable non-label indicators, not to infer causality or explain individual predictions.

The empirical gap lies in the limited integration of contextual readiness indicators, operational readiness classes, multiclass prediction, explicit leakage control, imbalance-sensitive evaluation, and ranking-stability analysis for traditional-market MSMEs, particularly in island-city settings. Accordingly, the study addresses three questions: (RQ1) how operational digital-readiness levels are distributed among traditional-market MSMEs in Ternate City; (RQ2) how a fixed untuned Random Forest baseline performs relative to a majority-class baseline; and (RQ3) which non-label indicators show the highest and most stable predictive importance. The contribution is integrative and empirical rather than algorithmic: the study establishes a transparent baseline against which later optimization can be assessed.

II. RESEARCH METHODOLOGY

A. Study Design, Setting, and Instrument

The study combined a cross-sectional survey with machine-learning classification. The analysis was descriptive and predictive rather than causal. It included 159 of 373 identified traders operating in Gamalama Market, Higienis Market, and Bastiong Market in Ternate City. One respondent represented each stall or business unit. Participants were recruited through non-probability convenience sampling based on presence, availability, and willingness to participate; consequently, the results are interpreted as an empirical baseline rather than population estimates for all traders in Ternate.

Data were collected from May 18 to July 31, 2026 by two enumerators using an assisted mixed-mode procedure that combined Google Forms and printed questionnaires. Respondents received information about the purpose of the study, confidentiality, voluntary participation, and the right to withdraw, and then provided verbal informed consent. Direct identifiers were excluded from the analytical dataset and each record was assigned a Respondent_ID.

B. Instrument and Operationalization of Digital Readiness

The questionnaire contained seven respondent and business characteristics (gender, education, age, business type, market location, business tenure, and monthly turnover) and 24 digital-readiness items (Q1-Q24) measured on a five-point ordinal scale. The items were organized into D1 Digital

Infrastructure and Access (Q1-Q4), D2 Technology-Use Competence (Q5-Q8), D3 Technology Implementation in Business Activities (Q9-Q12), D4 Digital Management and Understanding (Q13-Q16), D5 Support and Environment (Q17-Q20), and D6 Digital Trust and Risk Perceptions (Q21-Q24). Because formal content and construct validation had not yet been completed, Cronbach's alpha for the five score-construction items was used only as an exploratory indicator of internal consistency. The same exploratory check was applied to the 19 non-label predictor items. All items used behaviorally anchored response options ordered from the lowest to the highest level of the measured condition. For modeling, Q24 (fraud concern) was coded so that higher values indicate greater concern, whereas the other items were coded so that higher values indicate more favorable readiness conditions; Q24 was reverse-scored only for the reliability calculation. Because tree splits depend on the ordering of values rather than their direction, this coding does not change the information available to the Random Forest.

C. Class Construction and Leakage Control

The target variable was constructed from Q8, Q9, Q10, Q12, and Q14, which represent independent technology use and the implementation of technology in business activities [5]. Each item received equal weight, and the Digital Readiness Score (DRS) was calculated using Equation 1:

$$DRS_i = Q8_i + Q9_i + Q10_i + Q12_i + Q14_i \quad (1)$$

where DRS_i is the Digital Readiness Score of respondent i, and Q8_i, Q9_i, Q10_i, Q12_i, and Q14_i are that respondent's scores (1–5) on the five label items.

The theoretical DRS range is 5-25. A priori equal-width intervals classified scores of 5-11 as Low, 12-18 as Medium, and 19-25 as High. These thresholds are study-specific operational cutoffs, not validated universal readiness standards. For RQ1, frequencies and percentages were calculated for the three classes. For predictive modeling, Q8, Q9, Q10, Q12, Q14, the DRS, and the class label were removed from the predictor space to prevent direct target leakage. Conceptual similarity between the remaining predictors and the class-defining items is acknowledged as an interpretive limitation rather than direct leakage.

D. Data Preparation and Random Forest Baseline

Records were checked for duplicates, missing values, inconsistent categories, and values outside the valid scale. No observations were removed. After exclusion of the five label-construction items, the predictor space contained 19 ordinal readiness indicators and seven respondent/business characteristics, giving 26 predictors and 50 encoded features. Numeric missing values were handled by median imputation and categorical values by most-frequent imputation. The seven categorical variables were one-hot encoded with unknown categories ignored. All transformations were embedded in a Pipeline and ColumnTransformer so that preprocessing parameters were learned from the training portion of each validation fold.

Keeping preprocessing inside the validation pipeline was necessary to preserve the intended separation between training and validation data. Even simple transformations such as imputation and category encoding can transmit information across folds if they are fitted before cross-validation. The pipeline therefore applies the same sequence independently within every training fold and then transforms the corresponding validation fold using parameters learned only from the training data. This procedure does not remove all sources of modeling uncertainty, but it prevents a common form of preprocessing leakage and makes the reported baseline more reproducible.

Random Forest, originally formulated as an ensemble of randomized decision trees [27], was implemented as a fixed untuned baseline with 500 trees, Gini criterion, max_features='sqrt', bootstrap=True, random_state=42, and n_jobs=-1. No class weighting, SMOTE, over- or undersampling, feature selection, or hyperparameter search was applied. A DummyClassifier using the most_frequent strategy served as the majority-class comparator. Table 1 summarizes the sample characteristics, predictor structure, preprocessing steps, Random Forest configuration, and baseline restrictions used in the analysis.

Table 1. Summary of the Analytical Configuration for the Random Forest Baseline.

Component	Configuration
Sample	159 traders from three traditional markets in Ternate City
Predictors	19 ordinal readiness items + 7 respondent/business characteristics
Encoded features	50 after one-hot encoding
Preprocessing	Median/mode imputation; OneHotEncoder(handle_unknown="ignore")
Random Forest	500 trees; Gini; sqrt features; bootstrap; random_state=42
Baseline restriction	No balancing, feature selection, or hyperparameter tuning

E. Validation, Performance Metrics, and Predictive Importance

Performance was estimated using repeated stratified five-fold cross-validation with ten repetitions, resulting in 50 validation cycles. Stratification retained class proportions in each fold, while repetition reduced dependence on a single split. Accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1, weighted F1, per-class precision/recall/F1, Cohen's kappa, Matthews correlation coefficient (MCC), the confusion matrix, and one-vs-rest ROC-AUC were calculated. Balanced accuracy and macro-F1 were treated as primary metrics because they assign greater importance to performance across minority classes than raw accuracy alone.

Out-of-fold (OOF) predictions from the 50 cycles produced 1,590 prediction instances because each respondent was evaluated once in each repetition. These repeated predictions

were not interpreted as independent observations. For an additional respondent-level diagnostic, the ten hard-label OOF predictions for each respondent were consolidated by majority vote. ROC-AUC was computed from OOF probabilities and interpreted cautiously for the High class because only nine respondents belonged to that class.

The repeated-validation design was chosen to stabilize the baseline estimate without presenting repeated OOF records as an enlarged sample. Each repetition changes the assignment of respondents to training and validation folds, so the 50 cycles provide information about how sensitive performance is to resampling of the same 159 respondents. The aggregated confusion matrix is therefore used as a diagnostic summary of repeated predictions rather than as a basis for treating 1,590 records as statistically independent cases. This distinction is especially important for the High class: nine respondents can generate 90 OOF predictions across repetitions, but the underlying information still comes from only nine unique businesses.

RQ3 combined Gini importance, permutation importance, and ranking stability. Permutation importance was estimated from the change in macro-F1 across 20 permutations and, because the available computation used a model trained on the full dataset, was treated as exploratory complementary evidence rather than a fully out-of-sample importance estimate [28]. To address this limitation, permutation importance was additionally re-estimated out-of-fold: in each of the 50 validation cycles, the model fitted on the training folds was evaluated on its held-out fold, each predictor was permuted ten times, and the decrease in macro-F1 was averaged across cycles. Ranking stability was summarized by the mean rank and the proportion of the 50 validation cycles in which an indicator appeared in the top 10. The heuristic Composite Predictive Ranking Score (CPRS) was then calculated using Equation 2:

$$CPRS_j = 0.35\, G*_j + 0.35\, P*_j + 0.30\, S_j \quad (2)$$

where G*_j and P*_j are Gini and permutation importance normalized by their respective maximum values, and S_j is the top-10 appearance frequency divided by 50. CPRS is not a standardized explainable-machine-learning metric and does not indicate causal influence. It is used only to summarize relative importance and stability within this dataset. The weights were set heuristically rather than estimated. Gini and permutation importance received equal weight because they capture complementary aspects of relevance, namely impurity reduction within the trees and loss of predictive performance when a predictor is disrupted, and because each can be biased when predictors differ in scale of measurement or number of categories, a bias that is strongest for impurity-based importance [29]. Stability received a slightly lower weight because it reflects the consistency of a ranking rather than the magnitude of a predictor's contribution. Because this choice is not unique, a sensitivity analysis evaluated all 66 weight combinations on a 0.1 grid (weights summing to 1) for both the original components and a variant using out-of-fold permutation importance, and recorded how often each leading indicator remained in the top four and top five.

III. RESULTS AND DISCUSSION

A. Sample Profile and Readiness Distribution

All 159 observations passed the data-quality checks. The five items used for score construction produced a Cronbach's alpha of 0.7646, which indicates acceptable exploratory internal consistency but does not establish the unidimensionality or construct validity of the complete instrument. For the 19 non-label predictor items, Cronbach's alpha was 0.873, with corrected item-total correlations from 0.00 (Q1) to 0.69 (Q21). Domain-level values were acceptable for D2 (0.742), D4 (0.769), and D6 (0.749) but low for D1 (0.204) and D5 (0.401). Alpha reflects internal consistency, not validity or unidimensionality, and a single value computed across several distinct domains is of limited informative value, so the domain-level values are more relevant [30]. The low values for D1 and D5 are consistent with treating infrastructure and support items as formative conditions, for which internal consistency is not an appropriate quality criterion [31]. Because the Random Forest uses each item as a separate predictor, the feature-importance results are interpreted at the item level and do not depend on these domains forming unidimensional constructs. Women accounted for 57.23% of respondents; 57.86% had completed senior high school; and 57.86% traded at Gamalama Market. The largest age group was 36-45 years (33.96%), 57.86% had operated for more than six years, and the most common monthly turnover category was IDR 1-5 million (39.62%). These characteristics provide sample context but should not be interpreted as probability-based estimates of the full trader population.

The operational readiness classes were markedly imbalanced: 108 respondents (67.92%) were Low, 42 (26.42%) were Medium, and only 9 (5.66%) were High. This distribution directly affects the interpretation of model performance because a classifier can achieve relatively high accuracy by learning the dominant class. For that reason, balanced accuracy, macro-F1, class-specific metrics, and probability discrimination were considered alongside accuracy. This markedly imbalanced class distribution is illustrated in Figure 1.

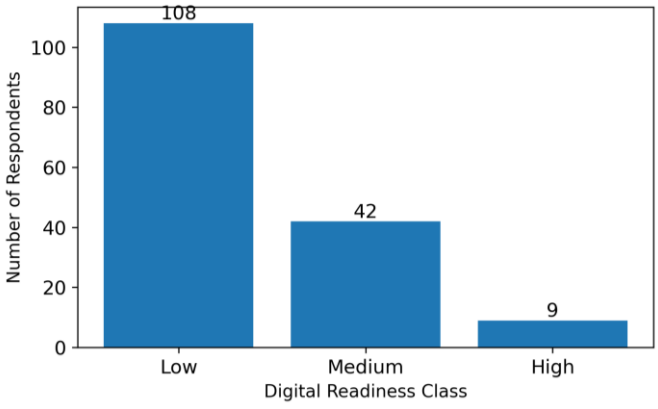


Figure 1. Distribution of Digital-readiness Classes Among the 159 Respondents.

B. Baseline Classification Performance and Class Imbalance

The majority-class DummyClassifier reached an accuracy of 0.6792, balanced accuracy of 0.3333, and macro-F1 of 0.2697. Random Forest improved these values to 0.7384, 0.4555, and 0.4470, respectively. Its weighted F1 was 0.7088, Cohen's kappa was 0.3739, and MCC was 0.3825. Relative to the majority baseline, the gains in accuracy, balanced accuracy, and macro-F1 were approximately 5.92, 12.22, and 17.73 percentage points. Thus, the non-label predictors contained information beyond the class prior, although the substantial difference between accuracy and balanced accuracy confirms that performance remained dominated by the Low class. Table 2 summarizes the overall performance of the Random Forest baseline relative to the majority-class comparator across the main evaluation metrics.

Table 2. Overall Performance of the Majority-class and Random Forest Baselines.

Metric	Dummy	Random Forest
Accuracy	0.6792	0.7384
Balanced accuracy	0.3333	0.4555
Macro-F1	0.2697	0.4470
Weighted F1	0.5495	0.7088
Cohen's kappa	0.0000	0.3739
MCC	0.0000	0.3825

The baseline was deliberately evaluated without class weighting or resampling. This design makes the impact of the original distribution visible rather than mixing data characteristics with the effect of a balancing intervention. The result is a useful benchmark: later experiments can measure whether class weighting, balanced Random Forest variants, or resampling actually improve minority-class recognition rather than merely changing overall accuracy [20]-[22].

The magnitude of the metric gap is also substantively important. An accuracy of 0.7384 could appear acceptable if considered alone, yet balanced accuracy of 0.4555 shows that average class recall remained low once each readiness level received equal weight. Macro-F1 leads to the same conclusion because it gives the three classes equal influence rather than allowing the 108 Low respondents to dominate the score. Cohen's kappa and MCC were only moderate, further indicating that the baseline learned useful structure without producing consistently reliable multiclass assignments. This combination of metrics is therefore more informative than a single accuracy value for an imbalanced readiness problem.

C. Per-Class Errors and Probability Discrimination

Class-level results reveal the main limitation of the baseline. The Low class achieved precision of 0.8056, recall of 0.9093, and F1 of 0.8543. The Medium class reached 0.5203, 0.4571, and 0.4867, while all three metrics for the High class were 0 because no High observation was correctly assigned the High label. Across the 1,590 aggregated OOF predictions, 982 Low and 192 Medium instances were correctly classified. All 90 High-class instances were misclassified: 79 as Medium and 11 as Low. Most errors therefore occurred between adjacent readiness levels, suggesting that the decision boundaries between neighboring classes were not sufficiently distinct. The complete pattern of classification errors is visualized in Figure 2, while the corresponding class-specific performance metrics are summarized in Table 3.

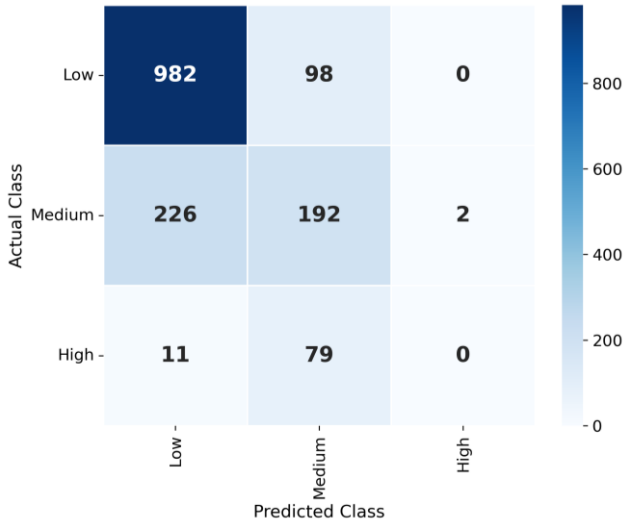


Figure 2. Aggregated Confusion Matrix for the Random Forest Baseline Based on 1,590 Out-of-fold Predictions (159 Respondents × 10 Repetitions; Not Independent Cases).

Table 3. Per-class Performance From Aggregated OOF Predictions

Class	Precision	Recall	F1	Support
Low	0.8056	0.9093	0.8543	1,080
Medium	0.5203	0.4571	0.4867	420
High	0.0000	0.0000	0.0000	90

The probability-based evaluation provides a complementary view. One-vs-rest AUC values were 0.876 for Low, 0.784 for Medium, and 0.859 for High, with a macro-AUC of 0.840. These class-wise one-vs-rest AUC values are shown in Figure 3. The High-class AUC of 0.859 is not inconsistent with recall of 0. ROC-AUC evaluates whether the model tends to rank High observations above non-High observations across thresholds, whereas recall is based on the final multiclass label chosen by the highest predicted probability. The baseline therefore captured a probabilistic High-class signal, but that signal was not strong enough to win the final class decision. This pattern supports later evaluation of class-balancing strategies and decision-threshold adjustment.

The respondent-level majority-vote diagnostic produced 116 correct and 43 incorrect classifications (accuracy 0.7296). Final predictions assigned 123 respondents to Low and 36 to Medium; none were assigned High. The errors comprised 24 Medium respondents classified as Low, 10 Low as Medium, eight High as Medium, and one High as Low. This again shows

that the main challenge is not only minority recognition but also separation of adjacent readiness levels.

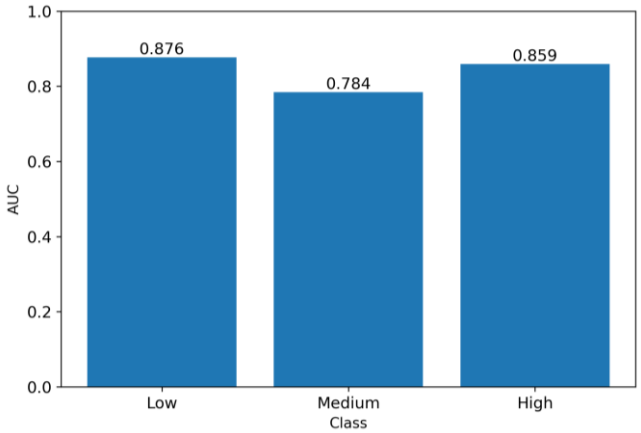


Figure 3. One-vs-rest ROC-AUC Values for the Low, Medium, and High Classes.

The direction of these errors is relevant because the readiness labels have a natural order. Most mistakes moved respondents by one level, particularly Medium to Low and High to Medium, rather than producing extreme Low-to-High or High-to-Low assignments. This suggests that the predictor profiles contain an ordered signal, but the nominal multiclass Random Forest does not explicitly model that order. The pattern supports future comparison with ordinal classification and cost structures that penalize larger readiness-distance errors more heavily than adjacent-class errors.

D. Predictive Indicators and Their Interpretation

The most informative indicators were concentrated in digital knowledge, perceived business value, and device adequacy. Q13 (online marketing knowledge) had the highest Gini importance (0.1021), followed by Q15 (technology importance in business; 0.0813), Q16 (perceived benefits of technology for sales; 0.0683), and Q7 (online selling knowledge; 0.0666). These four indicators appeared in the top 10 during all 50 validation cycles. Q2 (smartphone adequacy) appeared in the top 10 in 92% of cycles. CPRS ranked the same five indicators at the top, as summarized in Table 4. The broader ranking of the ten most informative non-label indicators according to CPRS is presented in Figure 4.

Permutation importance did not reproduce the Gini ranking exactly. Q3 had the largest permutation value (0.0224), followed by Q13 (0.0190), Q15 (0.0090), Q24 (0.0073), and Q16 (0.0072). The difference is expected because impurity-based importance measures how much a feature contributes to tree splits, whereas permutation importance measures the deterioration in macro-F1 when that feature is disrupted. Using both measures together with stability therefore provides a more cautious interpretation than relying on one ranking alone, especially under class imbalance [23].

Table 4. Five Leading Non-label Indicators by CPRS

Rank	Indicator	Gini	Mean rank	Top-10	CPRS
1	Q13 - Online marketing knowledge	0.1021	1.32	100%	0.9466
2	Q15 - Technology importance in business	0.0813	2.12	100%	0.7194
3	Q16 - Technology benefits for sales	0.0683	3.22	100%	0.6464
4	Q7 - Online selling knowledge	0.0666	3.46	100%	0.5325
5	Q2 - Smartphone adequacy	0.0374	7.44	92%	0.4475

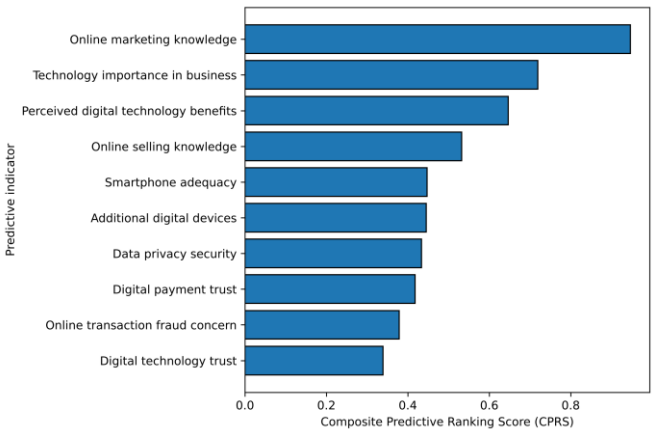


Figure 4. Top Ten Non-label Predictive Indicators Ranked by the Composite Predictive Ranking Score (CPRS).

Ranking stability adds information that a single fitted-model importance score cannot provide. Q13, Q15, Q16, and Q7 remained in the top 10 across all 50 validation cycles, indicating that their prominence was not tied to one particular partition of the limited sample. Q2 was slightly less stable but still appeared in 92% of cycles. This repeated appearance does not convert feature importance into causal evidence; it only strengthens the conclusion that these non-label indicators repeatedly contributed to the baseline's predictive structure under different training-fold compositions.

The robustness of these rankings was examined in two additional analyses, summarized in Table 5. First, out-of-fold permutation importance confirmed the leading indicators: Q15 (0.0459 ± 0.0380), Q13 (0.0422 ± 0.0381), Q7 (0.0243 ± 0.0257), and Q16 (0.0165 ± 0.0333) produced the four largest mean decreases in macro-F1 among the 26 predictors, with positive values in 92%, 90%, 84%, and 64% of validation

cycles, respectively. Their large standard deviations reflect the small validation folds of about 32 respondents. When the in-sample permutation component of CPRS was replaced by these out-of-fold values, the same four indicators occupied the first four positions. Q7, whose in-sample permutation value was close to zero, moved up to third position, indicating that its earlier low permutation score reflected evaluation on the training data rather than a lack of predictive information.

Second, the weight sensitivity analysis showed that the leading set does not depend on the chosen weights. With the original components, Q13 and Q15 remained in the top four under all 66 weight combinations and Q16 under 98.5% of them, and equal weights (1/3 each) reproduced the original top five exactly. Q7 remained in the top four in 62.1% of combinations and dropped out mainly when the in-sample permutation component was weighted heavily or the stability component lightly. With out-of-fold permutation, Q13, Q15, Q7, and Q16 formed the top four under every combination. In contrast, Q2 was weight-sensitive: it appeared in the top five in only 19.7% of combinations with the original components, ranked seventh in the out-of-fold variant, and had a slightly negative out-of-fold permutation value (−0.0028). Q2 therefore contributed consistently to tree splits, as shown by its 92% top-10 frequency, but its predictive information appears largely shared with correlated predictors; it is better interpreted as a supporting than as a leading indicator.

Table 5. Robustness of the Leading Indicators to Out-of-fold (OOF) Permutation Importance and CPRS Weighting

Indicator	OOF permutation (mean ± SD)	CPRS rank (original / OOF)	Top-4 share, original (%)	Top-4 share, OOF (%)
Q13	0.0422 ± 0.0381	1 / 1	100.0	100.0
Q15	0.0459 ± 0.0380	2 / 2	100.0	100.0
Q16	0.0165 ± 0.0333	3 / 4	98.5	100.0
Q7	0.0243 ± 0.0257	4 / 3	62.1	100.0
Q2	−0.0028 ± 0.0189	5 / 7	1.5	0.0

Note: Top-4 share is the percentage of 66 weight combinations (step 0.1, weights summing to 1) in which the indicator ranked in the top four; OOF permutation values are averaged over 50 validation cycles.

Conceptually, Q13, Q15, and Q16 belong to D4 Digital Management and Understanding, Q7 to D2 Technology-Use Competence, and Q2 to D1 Digital Infrastructure and Access. Indicators related to trust and risk also appeared in subsequent positions. The pattern supports the multidimensional understanding of digital readiness described in earlier work [3]-[6]: device availability matters, but knowledge of how technology can support marketing and sales and the perceived value of that technology provide stronger predictive separation in this dataset. These are associations with the model's class predictions, not evidence that changing any single indicator will causally increase readiness.

The respondent profiles reinforce this interpretation. Mean Q7 scores were 2.20, 3.43, and 3.44 for the Low, Medium, and High classes, indicating little separation between Medium and High. Q13 showed a clearer gradient of 1.71, 3.00, and 3.67; similar gradients appeared for Q15 (2.49, 3.69, 4.11) and Q16 (2.30, 3.40, 4.00). The model's difficulty distinguishing Medium from High is therefore consistent with overlap in some non-label indicators, while other indicators provide more stable ordering across the readiness levels.

E. Contribution and Practical Implications

The main contribution is the connection between descriptive digital-readiness assessment and predictive evaluation. Earlier studies provide multidimensional readiness concepts and evidence on determinants of adoption [3]-[10], whereas this analysis asks whether broader non-label indicators can distinguish operationally defined readiness classes. The separation of Q8, Q9, Q10, Q12, and Q14 from the predictor space is a methodological safeguard because it prevents the Random Forest from directly reusing the items that define the class. Combined with a majority-class comparator, repeated stratified validation, imbalance-sensitive metrics, and stability analysis, this creates a transparent baseline for later improvement.

This predictive perspective also clarifies what the model does and does not measure. The outcome is an operational readiness class derived from a study-specific scoring rule; it is not a direct measure of business success, digital maturity, or completed digital transformation. Readiness describes an initial condition for digital change [3], while the subsequent transformation process depends on organizational capabilities, resources, and external support [12]-[16]. Consequently, the leading predictors should be interpreted as signals that distinguish the present readiness classes, not as proof that the corresponding businesses have successfully transformed their operations.

The leakage-control procedure is particularly important for this distinction. If the five items used to construct the DRS had remained among the predictors, high performance could have reflected little more than reconstruction of the scoring rule. Removing those items forces the model to rely on broader readiness information that was not directly used to define the target. The remaining conceptual proximity among indicators cannot be eliminated completely, but the design makes the source of predictive information more transparent and gives the feature rankings a clearer diagnostic meaning.

The indicator rankings also have practical value for program design. Support for Low-readiness traders should not be limited to devices or Internet access; it should include basic online marketing, social-media and WhatsApp use for business, digital payments, and transaction security. Medium-readiness traders may need assistance in using digital tools more consistently, integrating promotion with transaction processes, and managing product information. For more

advanced traders, support can emphasize digital records, customer management, transaction data, and multichannel sales. A tiered approach is consistent with the view that MSME digitalization is gradual and should begin from existing capabilities and constraints [16].

Program use should nevertheless remain group-oriented rather than automated at the individual level. The baseline does not recognize the High class reliably, and the operational cutoffs have not been externally validated. A more defensible use at this stage is to inform the design of training modules and identify topics that deserve attention across trader groups. For example, the strong ranking of online marketing knowledge and perceived technology benefits suggests that capability-building should connect technical demonstrations with concrete business use cases, while smartphone adequacy and security-related indicators show that access and safe-use conditions should remain part of the intervention design.

Trust and risk perceptions also deserve attention. Data privacy, confidence in digital payments, fraud concerns, and perceived security were among the higher-ranked indicators after the top five. Training can therefore combine technical skills with practical safeguards such as password and PIN protection, transaction verification, fraud-scheme awareness, and responsible handling of digital footprints. Collaboration among market managers, local government, universities, financial-service providers, and MSME communities may help align training with local needs. The current Random Forest should remain a diagnostic research tool; it is not sufficiently balanced or validated to determine the readiness level of an individual trader.

From a monitoring perspective, the same indicators can also guide what should be measured after an intervention. If a training program targets online marketing, digital payments, or transaction security, follow-up assessment should examine whether knowledge and routine business use improve, not only whether participants report owning devices or attending training. Repeated measurement of the operational indicators could reveal whether traders move from basic familiarity toward consistent integration of technology. Such monitoring would also provide richer data for future predictive models than a one-time cross-sectional snapshot.

F. Limitations and Future Model Development

Several limitations constrain interpretation. Convenience sampling and the unequal number of respondents across markets limit generalizability. The six domains are literature-derived groupings because the instrument has not undergone formal content and construct validation. Cronbach's alpha of 0.7646 applies only to the five score-construction items and does not establish validity of the full instrument. The overall alpha of 0.873 for the 19 predictor items does not indicate unidimensionality, and the low consistency of D1 and D5 confirms that the domains are organizing categories rather than validated latent constructs; expert content-validity review and formal measurement modeling remain necessary before domain-level conclusions are drawn. The Low, Medium, and High cutoffs were defined a priori from theoretical score intervals and should not be treated as a universal standard. The High class contained only nine respondents, severely limiting the examples available for learning its decision boundary. Under five-fold stratification, each training fold contained only about seven High-class respondents, and the 90 High-class OOF predictions in Figure 2 reflect the same nine businesses rather than additional evidence. Together with convenience sampling, this means that the present results mainly describe the separation between Low and Medium readiness and should not be used to characterize high-readiness traders. Future data collection should deliberately oversample digitally active traders, for example through stratified or purposive recruitment, so that each readiness class contains enough cases for stable learning and class-specific evaluation.

The feature-importance analysis also has limitations. Permutation importance in the original CPRS was computed from a model fitted to the full dataset; the out-of-fold re-estimation confirmed the same four leading indicators, but its values varied considerably across cycles because each validation fold contained only about 32 respondents. CPRS is a heuristic combination of Gini importance, permutation importance, and ranking stability; it is not a standardized explainability metric. The sensitivity analysis showed that the leading set is robust to the choice of weights, but lower-ranked positions, including the fifth position, should be interpreted cautiously. In addition, one-hot-encoded categorical features and raw ordinal indicators do not have identical granularity, so the interpretation intentionally focuses on non-label ordinal indicators that remained informative across methods.

The cross-sectional design imposes an additional limitation. Predictor values and readiness classes were observed during the same study period, so the analysis cannot establish temporal precedence or causal effects. Business conditions, connectivity, technology costs, and trader practices may also change over time. Repeated measurement would be useful for testing whether the same indicators predict movement between readiness levels rather than merely differentiating respondents at one point in time. Such longitudinal evidence would also help determine whether the operational score thresholds remain meaningful as digital practices evolve.

Future work should retain the present Random Forest as a fixed comparator while addressing imbalance through class weighting, balanced Random Forest, cost-sensitive learning, or resampling performed only within each training fold [20]-[22]. Beyond imbalance handling, subsequent studies may investigate metaheuristic optimization approaches, such as Particle Swarm Optimization (PSO), for hyperparameter optimization and feature subset selection to improve minority-class recognition while maintaining model interpretability. Because the outcome is ordinal, ordinal classification should be compared with nominal multiclass modeling. The strong High-class AUC despite zero recall also motivates threshold tuning, provided that thresholds are selected from training data rather than optimized on held-out observations. Hyperparameter tuning and feature selection should be nested within cross-validation to reduce selection bias.

Interpretability should then move beyond global rankings. Explainable methods such as SHAP can describe the direction

and magnitude of feature contributions [25], while intervention-oriented analyses can connect prediction results with organizational change needs [26]. External validation across other markets and regions, a larger and more balanced sample, and formal validation of the questionnaire are necessary before a classifier is used in decision support. The objective of subsequent development should be not only higher predictive scores but also balanced class recognition, stable explanations, calibrated probabilities, and practical relevance to MSME digitalization programs.

A staged validation strategy would be preferable. First, the questionnaire should undergo expert review and construct validation on a larger sample. Second, model development should compare imbalance treatments and ordinal alternatives using nested validation, with calibration and High-class recall reported explicitly. Third, the selected model should be tested on traders from markets not used during development. Only after these steps should individual-level decision support be considered, and even then the model output should complement field assessment rather than replace it. This sequence preserves the present study's role as a baseline while defining a clear path toward a more reliable and interpretable system.

IV. CONCLUSION

This study established a transparent Random Forest baseline for classifying operational digital-readiness levels among 159 traditional-market MSME respondents in Ternate City. The outcome distribution was strongly imbalanced: 67.92% Low, 26.42% Medium, and 5.66% High. Random Forest outperformed the majority-class comparator, increasing macro-F1 from 0.2697 to 0.4470 and balanced accuracy from 0.3333 to 0.4555, but it produced no true positives for the High class. The High-class AUC of 0.859 indicates that a probability signal existed even though the final multiclass decisions remained biased toward lower readiness levels. Q13, Q15, Q16, and Q7 were the most informative and stable non-label indicators and remained the leading set under out-of-fold permutation importance and all 66 tested CPRS weightings, emphasizing the predictive relevance of digital knowledge, perceived business value, and online selling competence; Q2 (device adequacy) also ranked highly but was sensitive to the weighting scheme. These findings do not imply causality and the model is not suitable for individual readiness decisions. Its value lies in providing a leakage-controlled, imbalance-aware benchmark for future work on class balancing, ordinal classification, probability calibration, feature selection, and explainable machine learning.

CONFLICT OF INTEREST

The authors declare no conflict of interest involving the authors, participants, institutions, or organizations associated with this study.

ACKNOWLEDGEMENT

The authors acknowledge the Faculty of Engineering, Universitas Khairun, for supporting this work through the Penelitian Kompetitif Unggulan Perguruan Tinggi (PKUPT) research grant scheme.

Consensus was used to locate potentially relevant scholarly literature. ChatGPT (OpenAI) supported brainstorming, literature synthesis, manuscript organization, academic-language refinement, and Indonesian-English translation. The authors reviewed and verified all AI-assisted material against the underlying sources where applicable, revised the outputs, and accept full responsibility for the accuracy, integrity, interpretation, and final content.

REFERENCES

[1] L. Anatan and Nur, "Micro, Small, and Medium Enterprises' Readiness for Digital Transformation in Indonesia," *Economies*, vol. 11, no. 6, p. 156, 2023, doi: 10.3390/economies11060156.

[2] J. Díaz-Arancibia, J. Hochstetter-Diez, A. Bustamante-Mora, S. Sepúlveda-Cuevas, I. Albayay, and J. Arango-López, "Navigating digital transformation and technology adoption: A literature review from small and medium-sized enterprises in developing countries," *Sustainability*, vol. 16, no. 14, p. 5946, 2024, doi: 10.3390/su16145946.

[3] R. Silva, H. S. Mamede, and V. Santos, "Clarification of the present understanding of the assessment of an organization's digital readiness in SMEs," *Emerging Science Journal*, vol. 7, no. 6, pp. 2279–2307, 2023, doi: 10.28991/ESJ-2023-07-06-025.

[4] K. K. Rafiah, S. Widianto, I. Kamal, A. Shofiana, A. M. Fajar, and A. A. Rudini, "Digital readiness of SMEs: An insight from Indonesia," *AFEBI Management and Business Review*, vol. 7, no. 1, pp. 12–26, 2022, doi: 10.47312/ambr.v7i01.517.

[5] S. R. Pingali, S. Singha, S. Arunachalam, and K. Pedada, "Digital readiness of small and medium enterprises in emerging markets: The construct, propositions, measurement, and implications," *J. Bus. Res.*, vol. 164, p. 113973, 2023, doi: 10.1016/j.jbusres.2023.113973.

[6] I. Motjolopane and M. Chanza, "Digital transformation dimensions for evaluating SMEs' readiness for big data analytics and artificial intelligence: A review," *International Journal of Research in Business and Social Science*, vol. 12, no. 7, pp. 583–595, 2023, doi: 10.20525/ijrbs.v12i7.2837.

[7] F. Kurniasari, E. D. Lestari, and H. Tannady, "Pursuing long-term business performance: Investigating the effects of financial and technological factors on digital adoption to leverage SME performance and business sustainability—Evidence from Indonesian SMEs in the traditional market," *Sustainability*, vol. 15, no. 16, p. 12668, 2023, doi: 10.3390/su151612668.

[8] D. C. Guritno, A. N. Widya, A. Juwita, and R. Aryo, "The role of digitalization in the business turnover of women-

owned textile MSMEs in traditional markets: Evidence from Klewer Market, Surakarta," *Optimum: Jurnal Ekonomi dan Pembangunan*, vol. 16, no. 1, pp. 44–54, 2026, doi: 10.12928/optimum.v16i1.14563.

[9] R. Nurhapsari and E. Sholihah, "Analysis of the factors of intention to use QRIS for MSMEs in Semarang City's traditional market," *Jurnal Ekonomi Modernisasi*, vol. 18, no. 2, pp. 199–211, 2022, doi: 10.21067/jem.v18i2.7291.

[10] M. Mujianto, H. Hartoyo, R. Nurmalina, and E. Z. Yusuf, "The unraveling loyalty model of traditional retail to suppliers for business sustainability in the digital transformation era: Insight from MSMEs in Indonesia," *Sustainability*, vol. 15, no. 3, p. 2827, 2023, doi: 10.3390/su15032827.

[11] A. Arief, M. Y. Abbas, I. H. A. Wahab, L. A. Latif, S. D. Abdullah, and D. I. Sensuse, "The smart islands vision: Towards smart city readiness in local government of archipelagos," *J. Phys. Conf. Ser.*, vol. 1569, p. 042006, 2020, doi: 10.1088/1742-6596/1569/4/042006.

[12] T. H. Nguyen, X. C. Le, and T. H. L. Vu, "An extended Technology-Organization-Environment (TOE) framework for online retailing utilization in digital transformation: Empirical evidence from Vietnam," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 4, p. 200, 2022, doi: 10.3390/joitmc8040200.

[13] V. A. Ta and C.-Y. Lin, "Exploring the determinants of digital transformation adoption for SMEs in an emerging economy," *Sustainability*, vol. 15, no. 9, p. 7093, 2023, doi: 10.3390/su15097093.

[14] N. Omrani, N. Rejeb, A. Maalaoui, M. Dabić, and S. Kraus, "Drivers of digital transformation in SMEs," *IEEE Trans. Eng. Manag.*, vol. 71, pp. 5030–5043, 2024, doi: 10.1109/TEM.2022.3215727.

[15] M. Kljajić Borštnar and A. Pucihar, "Multi-attribute assessment of digital maturity of SMEs," *Electronics (Basel).*, vol. 10, no. 8, p. 885, 2021, doi: 10.3390/electronics10080885.

[16] G. H. Sagala and D. Őri, "Toward SMEs digital transformation success: A systematic literature review," *Information Systems and e-Business Management*, vol. 22, pp. 667–719, 2024, doi: 10.1007/s10257-024-00682-2.

[17] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, pp. 685–695, 2021, doi: 10.1007/s12525-021-00475-2.

[18] Y. Herdiyeni, B. Juanda, N. Zhafira, L. Anggraeni, and M. Probokawuryan, "Machine-learning-based assessment and prediction of digital transformation performance for MSMEs in Indonesia," *Int. J. Inf. Technol. Decis. Mak.*, vol. 25, no. 1, pp. 339–367, 2026, doi: 10.1142/S0219622025500415.

[19] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.

[20] C. Kaope and Y. Pristyanto, "The effect of class imbalance handling on datasets toward classification algorithm performance," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[21] J. Dong and Q. Qian, "A density-based random forest for imbalanced data classification," *Future Internet*, vol. 14, no. 3, p. 90, 2022, doi: 10.3390/fi14030090.

[22] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels," *Technologies (Basel).*, vol. 13, no. 3, p. 88, 2025, doi: 10.3390/technologies13030088.

[23] U. Bradter, J. Altringham, W. Kunin, T. Thom, J. O'Connell, and T. Benton, "Variable ranking and selection with random forest for unbalanced data," *Environmental Data Science*, vol. 1, p. e30, 2022, doi: 10.1017/eds.2022.34.

[24] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, p. 52, 2020, doi: 10.1186/s40537-020-00327-4.

[25] C. Zhu, X. Liu, and D. Chen, "Prediction of digital transformation of manufacturing industry based on interpretable machine learning," *PLoS One*, vol. 19, no. 3, p. e0299147, 2024, doi: 10.1371/journal.pone.0299147.

[26] M. Zine, F. Harrou, M. Terbeche, and Y. Sun, "Evaluating e-learning readiness using explainable machine learning and key organizational change factors in higher education," *Educ. Inf. Technol. (Dordr).*, vol. 30, pp. 12905–12937, 2025, doi: 10.1007/s10639-025-13335-9.

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[28] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[29] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007, doi: 10.1186/1471-2105-8-25.

[30] K. S. Taber, "The use of Cronbach's alpha when developing and reporting research instruments in science education," *Res. Sci. Educ.*, vol. 48, no. 6, pp. 1273–1296, 2018, doi: 10.1007/s11165-016-9602-2.

[31] A. Diamantopoulos and H. M. Winklhofer, "Index construction with formative indicators: An alternative to scale development," *Journal of Marketing Research*, vol. 38, no. 2, pp. 269–277, 2001, doi: 10.1509/jmkr.38.2.269.18845.