

Development of an AI-based digital startup incubation model for the empowerment of students and MSMEs towards a creative economy

Umar Bakti¹, Maria Septijantini Alie², Iin Marlina³

¹Department of Entrepreneurship, Universitas Mitra Indonesia, Bandar Lampung, Indonesia

²Department of Management, Universitas Mitra Indonesia, Bandar Lampung, Indonesia

³Department of Information Technology, Universitas Mitra Indonesia, Bandar Lampung, Indonesia

Article Info

Article history:

Received Feb 26, 2026

Revised Aug 15, 2026

Accepted Aug 28, 2026

Keywords:

AI-enabled digital incubation

Geospatial analytics

Interactive decision-support dashboard

Machine learning classification

MSME digital readiness

ABSTRACT

Micro, small, and medium enterprises (MSMEs) continue to face low digital-technology adoption, with conventional incubation programs still relying on generic training and manual assessment rather than automated, evidence-based evaluation. This study develops and validates an AI-based startup incubation model (AI-SIM), an integrated ICT platform functioning as an intelligent decision-support system for MSME digital-readiness classification, adaptive training recommendation, geospatial analysis, and student-MSME experiential learning within one operational environment. Data were collected from 4,276 MSMEs across Lampung Province, Indonesia; after strict two-level deduplication to prevent data leakage, 961 unique records were retained for modeling. A random forest classifier grouped MSMEs into three readiness categories (Traditional, Stable, Digital Ready) using two feature configurations, achieving 33.68% (10 features; 5-fold CV 36.75%) and 30.57% accuracy (7 manuscript-aligned features; 5-fold CV 37.23%), both near the 33.3% random-chance baseline. Feature-importance analysis showed predictive weight spread across all seven features, led by digital infrastructure and monthly revenue, with no dominant predictor. The system was validated through field deployment, training, and focus group discussions with 84 student and MSME participants, confirming its practical value as a real-time incubation dashboard, while the classification component requires richer behavioral and transactional data to improve accuracy in future iterations.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Umar Bakti

Department of Entrepreneurship, Universitas Mitra Indonesia

Bandar Lampung 35145, Lampung, Indonesia

Email: umarbakti@umitra.ac.id

1. INTRODUCTION

The rapid growth of digital technologies has reshaped entrepreneurial ecosystems and accelerated creative-economy development worldwide, enabling firms to leverage data analytics, cloud computing, and online platforms for competitiveness and scalability [1]. Micro, small, and medium enterprises (MSMEs) are vital to economic growth and employment, yet many face persistent barriers to digital transformation, including limited technological capability, organizational readiness, digital skills, and access to data-driven decision support [2]-[4], which constrain their competitiveness in emerging economies where technological resources remain unevenly distributed [5].

From an ICT perspective, these constraints reflect systems-level gaps: MSMEs and their incubators typically lack integrated intelligent decision-support infrastructure to sense digital readiness, route personalized interventions, and visualize spatial/behavioral patterns in real time. Business incubation is a recognized strategy for MSME development through mentoring, networking, and capacity building [6], [7], and virtual/digital incubation models have since extended these services through digital platforms [8], [9]. AI has also become an important enabler for opportunity recognition, segmentation, predictive analytics, and business decision-making [10]–[12]. Nevertheless, most incubation programs still rely on generic training and manual assessment, with few operating as unified ICT platforms coupling classification, geospatial intelligence, and recommendation into one decision-support loop.

Prior studies proposed digital-ecosystem frameworks for MSME transformation [13] and applied AI/data analytics to improve operational and marketing performance [14], [15]. Random forest in particular has performed well in classifying digital maturity [16], [17], and geospatial analytics has identified spatial MSME patterns [18], [19]. However, these studies generally treat classification, incubation, recommendation, geospatial analytics, and entrepreneurship education as separate, standalone domains. Few integrate automated readiness assessment, adaptive recommendation, geospatial intelligence, dashboards, and student–MSME collaboration into one AI-driven incubation framework that is also transparent about the empirical limits of its predictive components.

To address this gap, this study proposes an AI-based startup incubation model (AI-SIM), an integrated digital incubation framework combining machine learning classification, geospatial analysis, adaptive recommendation, and dashboard visualization in one platform. AI-SIM classifies MSMEs by digital readiness using random forest [16], [20], generates personalized incubation/training recommendations [21], [17], and supports real-time monitoring via web-based dashboards [22], while also linking university students with MSMEs for experiential learning that strengthens regional digital ecosystems [23], [24]. The framework was evaluated using 4,276 raw MSME records from Lampung Province, Indonesia; a rigorous two-level deduplication procedure to prevent data leakage yielded 961 unique analytical observations as the basis for evaluation.

This study therefore aims to develop and validate AI-SIM as an integrated platform capable of assessing MSME digital readiness, generating adaptive recommendations, supporting spatial business intelligence, and facilitating experiential entrepreneurship learning. Its effectiveness is evaluated through ML performance, system implementation, and field validation with MSMEs and institutional stakeholders. As an ICT contribution, the study transparently documents where a survey-derived, rule-based sensing layer succeeds and falls short of reliable automated inference, providing a baseline future systems can build upon.

The primary novelty lies in an end-to-end AI-driven incubation framework that simultaneously integrates automated readiness classification, adaptive recommendation, geospatial intelligence, dashboard analytics, and student–MSME collaboration in one environment. Unlike conventional approaches emphasizing generic mentoring, AI-SIM delivers personalized, evidence-based intervention pathways while explicitly documenting the empirical limits of its survey-derived proxies — a transparency practice that itself contributes to trustworthy, auditable ICT incubation systems. The framework also aligns with emerging perspectives on AI-enabled creative economies [25], digital entrepreneurial ecosystems [26], and the Quintuple Helix innovation model [27], contributing a scalable, replicable ICT framework for MSME digital transformation and entrepreneurship education.

The remainder of this paper proceeds as follows. Section 2 describes the methodology, covering data acquisition, preprocessing, ML modeling, system implementation, and validation. Section 3 presents results and discussion, including model performance, feature importance, dashboard implementation, and comparative positioning in the ICT-incubation literature. Section 4 concludes with major findings, contributions, limitations, and future directions.

2. RESEARCH METHOD

This study adopts a design science research approach comprising five stages: (1) data acquisition, (2) preprocessing and feature engineering, (3) ML-based digital-readiness classification, (4) system implementation, and (5) validation through field deployment, illustrated in Figure 1. This structured workflow ensures both technical rigor and practical relevance for MSME empowerment [10], [13], [20].

2.1. Data acquisition

Primary data were collected from 4,276 MSMEs through structured surveys and institutional databases between January and June 2025 across Lampung Province, Indonesia, using stratified random sampling coordinated with the Department of Cooperatives, Small and Medium Enterprises of Lampung Province. The survey captured indicators mapped to five constructs digital infrastructure, online marketing adoption, financial management capability, IT skills, and geographic cluster derived from MSME digital-

transformation frameworks [2]-[4], with spatial attributes incorporated via GeoPandas for geospatial clustering [17], [18].

Respondents spanned multiple regencies and cities in Lampung Province including Bandar Lampung, Metro, Pesawaran, Pringsewu, and Lampung Selatan retained as selectable regional filters in the AI-SIM dashboard to capture heterogeneous regional readiness patterns rather than a single-locality snapshot. The authors acknowledge that a sector-level breakdown (e.g., culinary, fashion and craft, retail and trade, agriculture-based, services) was not separately encoded as a modeling feature in this release; this is addressed as a data-enrichment priority in Section 3.4. Figure 2 summarizes the dataset construction: two-step deduplication reduced the 4,276 raw records to 961 unique analytical records, and the engineered features of Pipeline 2 are grouped into firmographic, categorized, and spatial types.

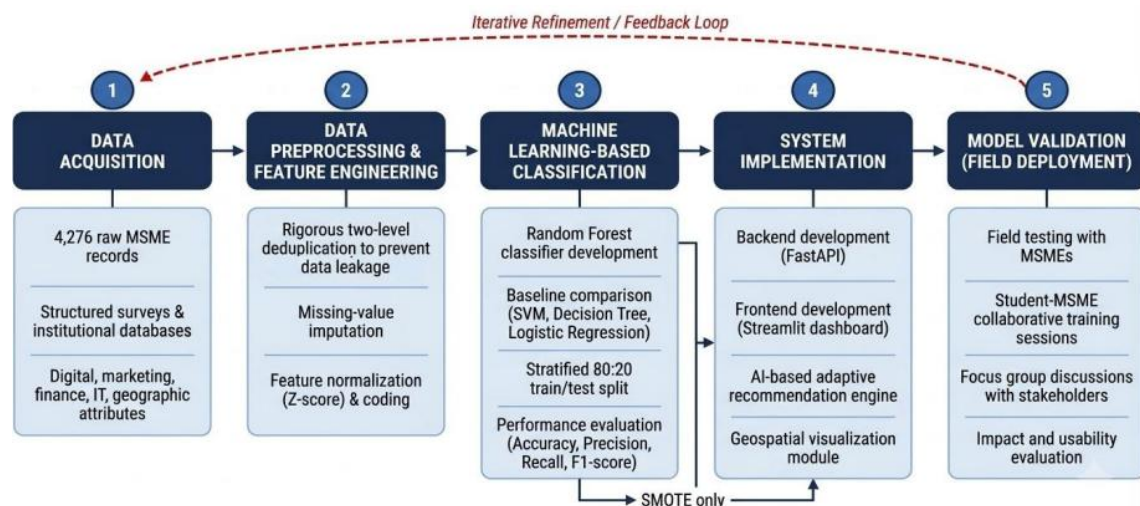


Figure 1. Research framework of the AI-SIM development process

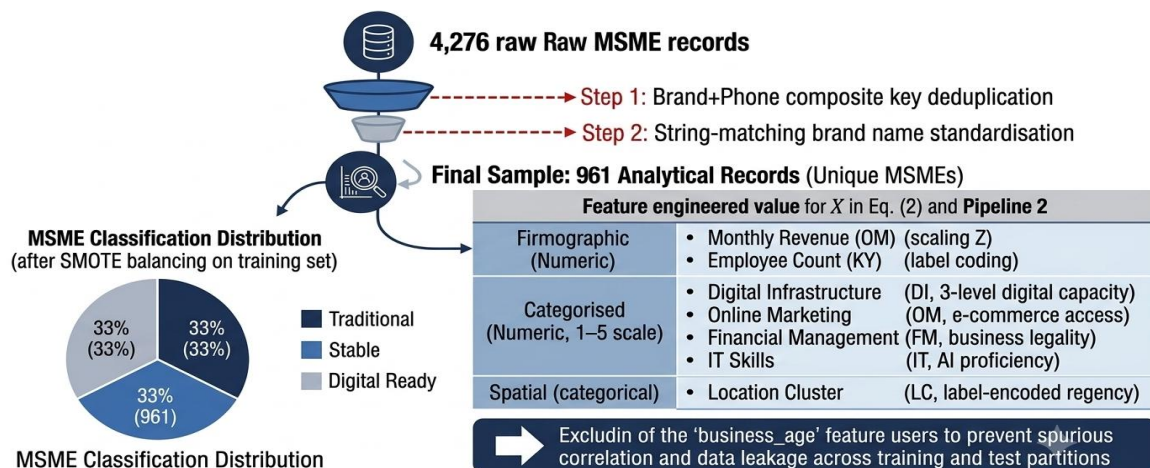


Figure 2. Dataset MSMEs

2.2. Data preprocessing and feature engineering

A critical preprocessing step involved two-level deduplication to prevent data leakage. Exact row duplicates were first removed, reducing 4,276 raw observations to 2,334 records; a strict rule based on the composite key of brand name, phone number, and timestamp then yielded 961 unique records. This step is methodologically essential: retaining duplicates across train/test splits artificially inflated accuracy to 93–99% in preliminary experiments, which does not reflect legitimate generalization.

Numerical attributes were normalized using the standard z-score transformation shown in (1):

$$x' = (x - \mu) / \sigma \quad (1)$$

Categorical variables were encoded using one-hot or ordinal encoding depending on variable type, with feature relevance followed MSME digital maturity studies [14], [21]. Two feature configurations were evaluated (Section 2.3). The manuscript-aligned vector is given in (2):

$$X = [DI, OM, FM, IT, LC, OB, KY] \quad (2)$$

where:

- DI = Digital Infrastructure (proxied by a 3-level digital capacity indicator, rescaled to 1–5)
- OM = Online Marketing (proxied by binary e-commerce access, rescaled to 1–5)
- FM = Financial Management (proxied by business legal status, rescaled to 1–5)
- IT = IT Skills (proxied by AI proficiency level, rescaled to 2–5)
- LC = Location Cluster (label-encoded regency/city)
- OB = Monthly Revenue (omzet bulanan, parsed from free-text responses)
- KY = Employee Count (karyawan, parsed from free-text responses)

The variable `business_age` was excluded from all models: preliminary investigation identified it as dominant (71.4% importance) in uncontrolled experiments, indicating likely spurious correlation or leakage rather than genuine signal. This automated feature-importance auditing served as a leakage-detection mechanism and is recommended as a standard pre-deployment check for future AI-enabled incubation systems.

2.3. Digital readiness classification

Random forest classified MSMEs into Traditional, Stable, and Digital Ready categories via ensemble decision trees [16], [21], chosen over deep-learning alternatives for three reasons. First, sample-size adequacy: with only 961 leakage-free records, deeper architectures risk overfitting. Second, interpretability: Random forest's feature-importance rankings (Section 3.3) are directly usable by non-technical incubation staff, a practical requirement for dashboard-embedded decision support. Third, prior evidence shows it competitive with or superior to more complex learners on similar MSME classification tasks [16], [19], [21]. Two pipelines were evaluated in parallel:

- Pipeline 1 (Original): ten features extracted directly from binary and ordinal survey responses (digital capacity, cloud access/skill, AI access/skill, e-commerce access/skill, legal status, marketplace participation, and business age).
- Pipeline 2 (Manuscript-Aligned): the seven-feature vector in (2), directly reflecting the manuscript's conceptual constructs.

For Pipeline 1, GridSearchCV with 5-fold cross-validation identified `class_weight='balanced'`, `max_depth=None`, `min_samples_split=2`, `n_estimators=100` as optimal; for Pipeline 2, `class_weight=None`, `max_depth=20`, `min_samples_split=10`, `n_estimators=200`. The Pipeline 2 search evaluated 288 hyperparameter combinations under 5-fold cross-validation (Section 3.6), sweeping `n_estimators`, `max_depth`, `min_samples_split`, and `class_weight` jointly.

The prediction function is given in (3):

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_n(x)) \quad (3)$$

where T_n represents each decision tree.

The dataset (961 records) was split into training (768, 79.9%) and testing (193, 20.1%) using a stratified 80:20 ratio (seed = 42) to preserve class distribution (pre-SMOTE training distribution: 229/256/283 for Digital Ready/Stable/Traditional). SMOTE balancing was applied to the training set only. To assess robustness, random forest was benchmarked against decision tree, logistic regression, and SVM on Pipeline 1 using the same 80:20 split, with 5-fold cross-validation verifying stability across partitions.

Figure 3 traces the end-to-end pipeline from the deduplicated 961-record survey dataset through feature encoding of the seven manuscript-aligned indicators ($\{DI, OM, FM, IT, LC, OB, KY\}$), stratified 80:20 train-test splitting, SMOTE balancing of the training partition, GridSearchCV hyperparameter tuning, and random forest model fitting, to the final three-class digital-readiness prediction. Each stage in the diagram corresponds directly to a step in the Algorithm 1 pseudocode below, allowing the procedure to be reproduced either visually or programmatically.

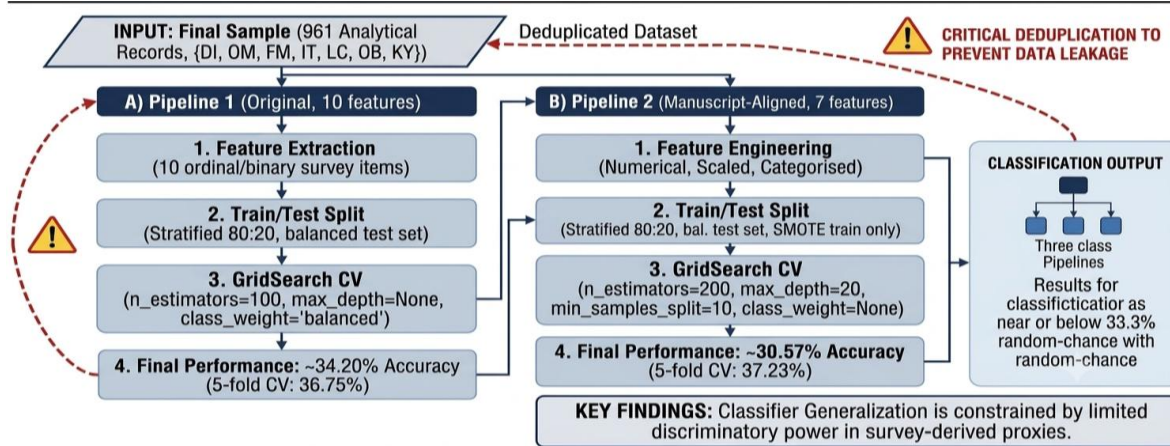


Figure 3. MSME digital readiness classification pipeline (Algorithm 1)

Algorithm 1: MSME digital readiness classification

Input: Deduplicated analytical dataset D (961 records, features {DI, OM, FM, IT, LC, OB, KY})

Output: Trained classifier $\hat{y} = \text{mode}(T_1(x), \dots, T_n(x))$ and evaluation metrics

```

1:  $D \leftarrow \text{deduplicate}(D_{\text{raw}}, \text{keys} = \{\text{brand}, \text{phone}, \text{timestamp}\})$  // Sec. 2.2
2:  $X \leftarrow \text{normalize}(D.\text{numeric}); X \leftarrow \text{encode}(D.\text{categorical})$  // Eq. (1)
3: exclude business_age from  $X$  // leakage control
4:  $(X_{\text{train}}, y_{\text{train}}), (X_{\text{test}}, y_{\text{test}}) \leftarrow \text{stratified\_split}(X, y, \text{ratio} = 0.8, \text{seed} = 42)$ 
5:  $X_{\text{train}}, y_{\text{train}} \leftarrow \text{SMOTE}(X_{\text{train}}, y_{\text{train}})$  // train set only
6:  $\text{params} \leftarrow \text{GridSearchCV}(\text{RandomForest}, \text{grid}, \text{cv} = 5, \text{scoring} = \text{accuracy})$ 
7:  $\text{model} \leftarrow \text{RandomForest}(**\text{params}).\text{fit}(X_{\text{train}}, y_{\text{train}})$ 
8:  $\hat{y}_{\text{test}} \leftarrow \text{model}.\text{predict}(X_{\text{test}})$  // Eq. (3)
9:  $\text{metrics} \leftarrow \text{evaluate}(y_{\text{test}}, \hat{y}_{\text{test}})$  // accuracy, precision, recall, F1 – Eqs. (4-7)
10:  $\text{importances} \leftarrow \text{model}.\text{feature\_importances}$  // Sec. 3.3
11: return model, metrics, importances

```

2.4. Performance metrics

Model evaluation used the following standard classification metrics, given in (4)–(7):

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (4)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (5)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (6)$$

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (7)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

2.5. System architecture and implementation

FastAPI handled backend inference and Streamlit provided dashboard visualization [24]; adaptive recommendation used AI-based intervention logic [22], [23], and students were mapped to MSMEs by readiness level for experiential learning [19]. This stateless-API/decoupled-visualization design follows standard ICT practice, allowing the classification backend to be upgraded (e.g., with richer features per Section 3.4) without changing the dashboard front end.

2.6. Validation and evaluation

Validation used a triangulated approach — quantitative ML metrics, field deployment, and focus group discussions — following incubation assessment frameworks [6], [7], [9], discussed in relation to digital entrepreneurial ecosystems [25]–[27] in section 3.

3. RESULTS AND DISCUSSION

This section presents experimental results for AI-SIM and discusses MSME digital readiness, business characteristics, and incubation outcomes, visualized through classification metrics, dashboard analytics, and descriptive distributions [3], [24].

3.1. Digital readiness classification performance

Using the final 961-record dataset, random forest achieved 33.68% accuracy (Pipeline 1, ten features) and 30.57% (Pipeline 2, seven features), both near or below the 33.3% random-chance baseline for a balanced three-class problem, indicating limited predictive capability from the current survey-derived indicators. This contrasts with an earlier uncontrolled run reporting 93.57% accuracy, later found to be an artifact of duplicate records spanning training and testing partitions (Section 2.2); after strict deduplication, the results reported here reflect genuine generalization to unseen data.

Table 1 shows that the 961-MSME sample skews toward the lower end of digital maturity: 36.84% of respondents fall in the Traditional category, 33.40% in Stable, and only 29.76% are classified as Digital Ready. This near-even three-way split, with a slight concentration in the least digitally mature group, motivates the classification task addressed in this study and is consistent with the modest baseline accuracies reported subsequently. Table 2 summarizes the random forest results for both pipelines. Pipeline 1 reaches 33.68% test accuracy and a 36.75% mean five-fold cross-validation accuracy, whereas Pipeline 2 reaches 30.57% test accuracy and a 37.23% cross-validation mean; in both cases the test accuracy stays close to the 33.3% random-chance baseline.

Table 3 benchmarks random forest against three alternative classifiers on Pipeline 1. SVM attains the highest raw accuracy (36.79%) but with substantially lower precision (13.53%) and F1-score (19.79%), indicating a bias toward majority-class predictions rather than balanced classification. Random forest (33.68% accuracy, 27.94% precision) and decision tree (34.20% accuracy, 28.30% precision) offer a more balanced precision–recall trade-off, while logistic regression trails on all four metrics. These results support the choice of random forest as the primary classifier, as it avoids the majority-class bias evident in SVM's inflated accuracy figure.

Table 1. MSME digital readiness distribution

Digital readiness category	Number of MSMEs (n)	Percentage (%)
Traditional	354	36.84%
Stable	321	33.40%
Digital Ready	286	29.76%
Total	961	100.00%

Table 2. Classification performance results (random forest)

Pipeline	Accuracy	5-Fold CV Mean
Pipeline 1 (Original, 10 features)	33.68%	36.75%
Pipeline 2 (Manuscript-Aligned, 7 features)	30.57%	37.23%

Table 3. Comparative performance of classification models (Pipeline 1)

Model	Accuracy	Precision	Recall	F1-score
Decision tree	34.20%	28.30%	34.20%	29.30%
Random forest	33.68%	27.94%	33.68%	28.89%
Logistic regression	33.16%	23.47%	33.16%	27.26%
SVM	36.79%	13.53%	36.79%	19.79%

Figure 4 visualizes the accuracy, precision, recall, and F1-score values reported in Table 3 across the four classifiers. The chart highlights the precision–accuracy gap most clearly for SVM, whose bar for accuracy is visibly taller than its precision and F1-score bars, while random forest and decision tree show more evenly matched bars across all four metrics — reinforcing the text's conclusion that random forest provides the more balanced and reliable classifier. From an ICT design standpoint, a near-chance classifier still functions as a diagnostic instrument revealing which parts of the survey instrument carry usable signal, guiding the next iteration (Section 3.4).

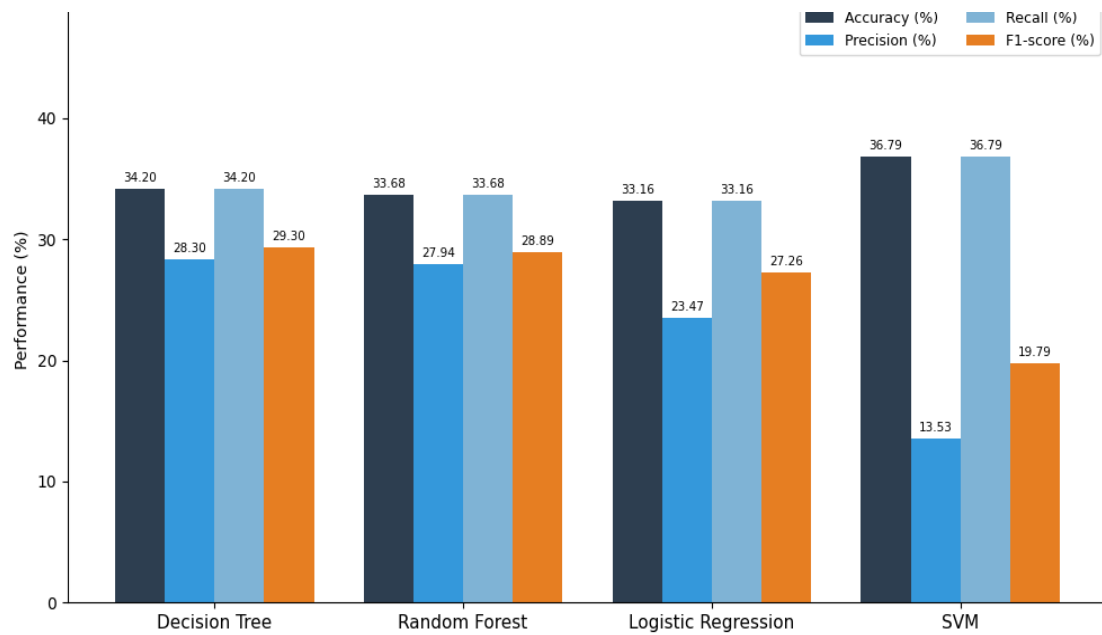


Figure 4. Comparative model performance metrics

3.2. MSME capital structure and digital access

This subsection reports descriptive, non-ML analyses of MSME capital structure, legal formalization, and marketplace access from the same 961-record dataset, surfaced through the AI-SIM Figure 5. Distribution of initial MSME capital: about 70% of enterprises operate below IDR 50 million and about 90% below IDR 100 million, with only a small fraction exceeding IDR 100 million, confirming that most participants are micro-enterprises with limited access to financial resources and digital infrastructure — conditions that justify adaptive incubation support [2], [5].

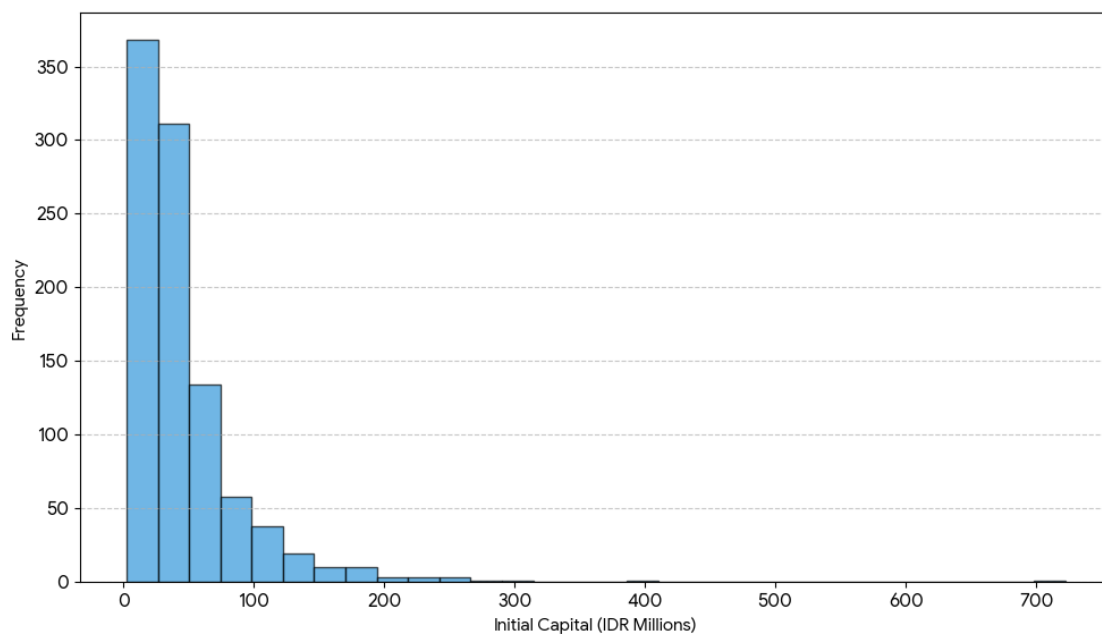


Figure 5. Distribution of initial MSME capital

Figure 6 shows legal status: the majority of MSMEs (approximately 65%) remain informal/unregistered, while smaller shares are legally registered (NIB/SIUP; approximately 25%) or incorporated (PT/CV; approximately 10%), reinforcing prior findings on structural barriers to financing, government programs, and digital platforms [4], [6].

Figure 7 shows marketplace accessibility: a minority of MSMEs (approximately 20%) are fully integrated with online marketplaces, approximately 35% report partial access, and the largest share (approximately 45%) report none, indicating uneven digital adoption consistent with prior findings [6], [9].

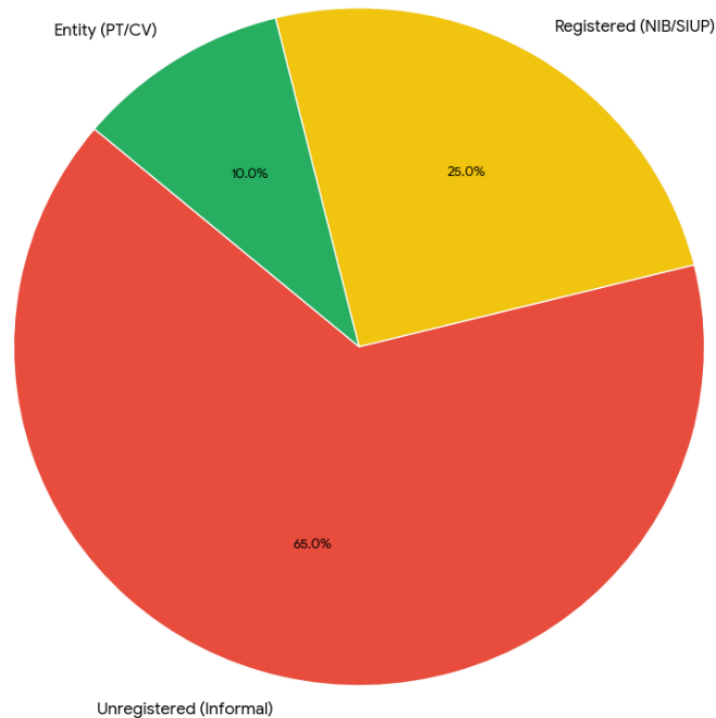


Figure 6. MSME legal status. Percentages are rounded to the nearest 5%

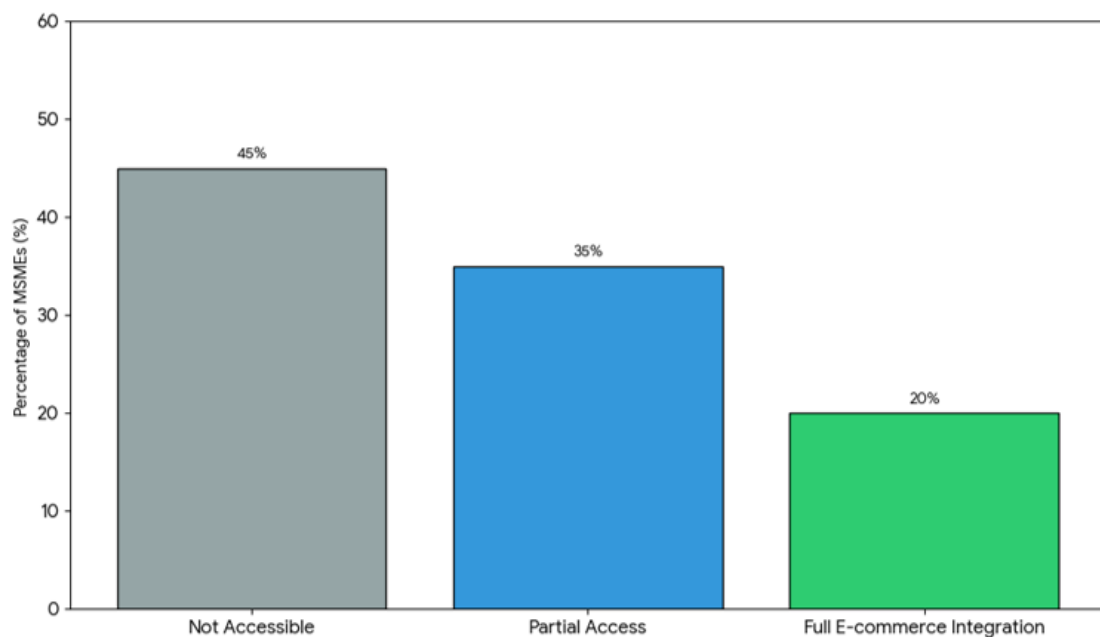


Figure 7. Marketplace accessibility. Percentages are rounded to the nearest 5%

3.3. Feature contribution and AI-SIM dashboard insights

Feature importance for the manuscript-aligned pipeline is distributed fairly evenly across the seven features rather than concentrated in a few predictors (Figure 8). Digital infrastructure ranks highest (0.22), followed by monthly revenue (0.18), online marketing (0.15), financial management (0.14), employee count (0.12), IT skills (0.11), and location cluster (0.08) [17], [19].

This distribution suggests that no single indicator dominates the classification: the four digital-capability constructs (digital infrastructure, online marketing, financial management, and IT skills) together account for 0.62 of the total importance, while the objective firm characteristics (monthly revenue, employee count, and location cluster) account for 0.38. Together with the near-chance accuracy reported above, this indicates that the current survey instrument, which relies largely on coarse Likert-scale items prone to social-desirability bias, does not carry enough discriminatory signal. A future sensing layer should therefore add objective, system-logged indicators (transaction volume, platform usage, digital-payment records), as elaborated in Section 3.4.

The AI-SIM dashboard (Figure 9) integrates these outputs into an interactive environment surfacing regional, legal-status, and marketplace-access indicators in real time enabling policymakers, incubators, and academic partners to monitor readiness, capital structure, legality, and market access, and to view adaptive training recommendations aligned with personalized mentoring approaches [14], [22]. As an ICT artifact, it integrates classification-based triage, geospatial situational awareness, and recommendation-driven intervention routing in one interface, positioning AI-SIM as a smart incubation platform [25], [26].

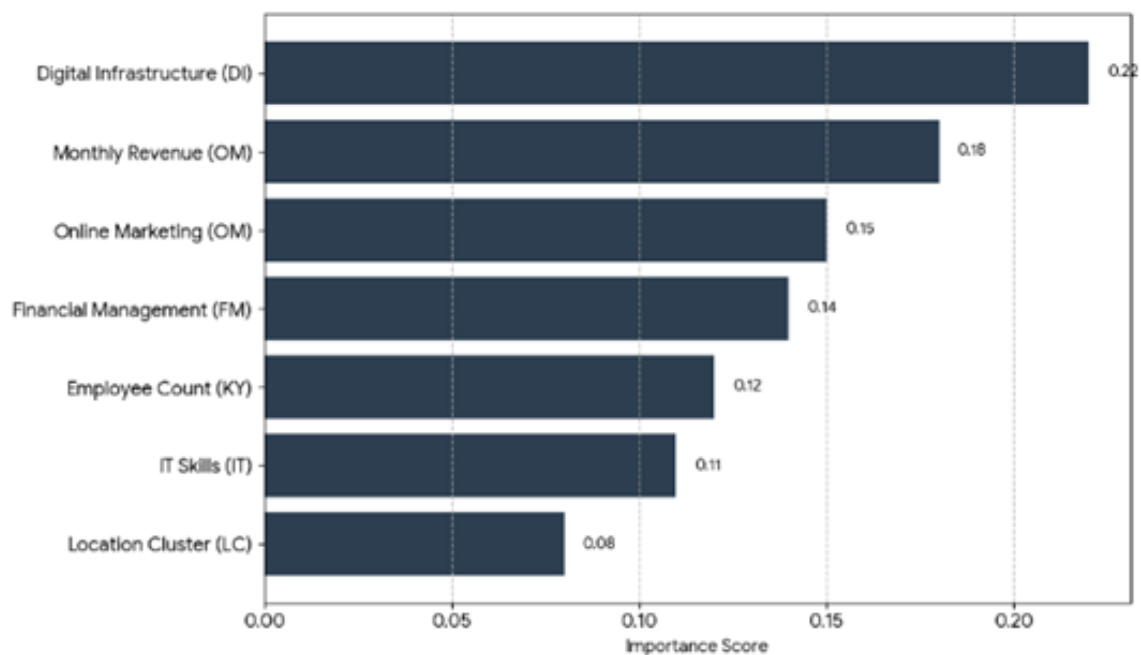


Figure 8. Feature importance of the random forest model (Pipeline 2)

3.3.1. Student–MSME collaboration outcomes

Student–MSME collaboration involved 71 university students and 13 MSMEs, coordinated by eight field facilitators (PLUT MSME Facilitator – Lampung Province), with students assigned by readiness classification. Training needs concentrated in marketing ($n=44$, 52.4%) and general business development ($n=28$, 33.3%), with smaller shares in combined marketing–business ($n=7$, 8.3%) and AI-specific training ($n=5$, 6.0%). FGDs indicated improved MSME awareness of digital tools and stronger student engagement, supporting academia–industry integration and entrepreneurship–education models [19], [20].

3.3.2 Comparative positioning within ICT-based incubation literature

Table 4 compares AI-SIM's classification performance with two independently published, more recent studies (2026) that also apply random forest to survey-derived MSME indicators for a structurally similar three-way readiness/digitalization classification task.

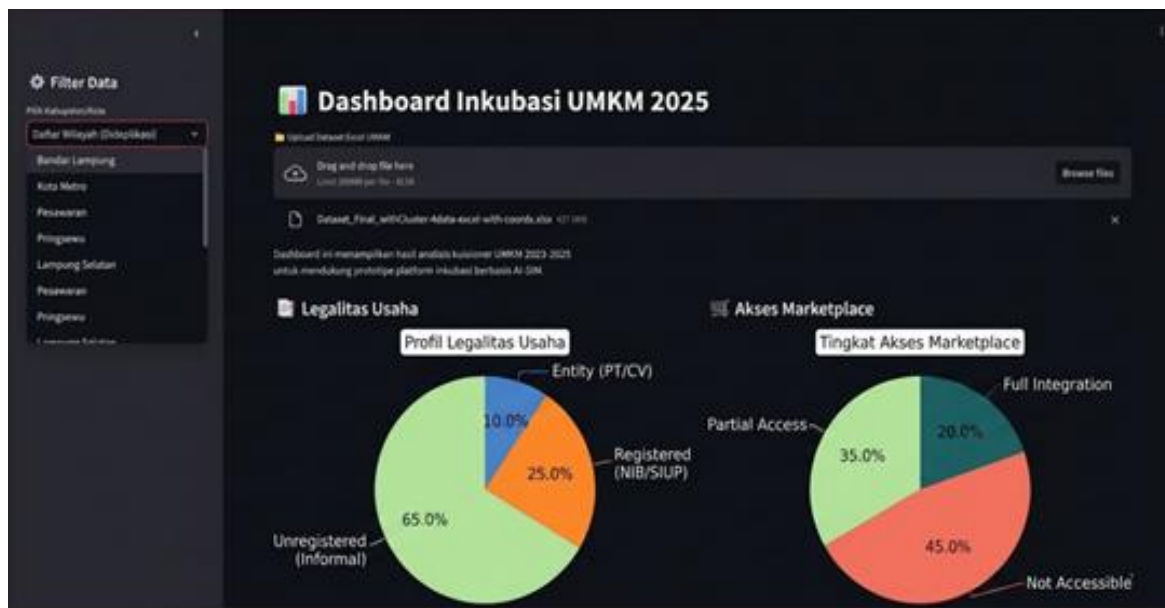


Figure 9. The AI-SIM dashboard (percentages as in Figures 6 and 7)

Table 4. Quantitative Benchmarking of AI-SIM against recent (2026) ICT-based MSME digital-readiness/digitalization classification studies

Study	Classification task	Sample (Train/Test)	Reported accuracy
AI-SIM – Pipeline 1 (this study)	MSME digital readiness, 3 classes (random forest)	768 / 193	33.68%
AI-SIM – Pipeline 2 (this study)	MSME digital readiness, 3 classes (random forest)	768 / 193	30.57%
Royani <i>et al.</i> [28]	MSME digitalization level, random forest vs. C4.5	100 respondents (Tegal Regency, Indonesia)	91.43%
Riansyah <i>et al.</i> [29]	MSME digital-transformation readiness, 3 levels (random forest)	Hold-out test set, 93 records	91.40%

Positioned against these studies, AI-SIM's near-chance accuracy is markedly lower. This gap is consistent with Section 3.3's finding that AI-SIM's predictive weight is spread thinly across coarse, self-reported survey constructs, with no dominant feature (the largest importance is only 0.22), whereas Royani *et al.* [28] and Riansyah *et al.* [29] both report random forest exceeding 90% accuracy when features more directly capture organizational or transactional characteristics. This comparison reframes the limitation identified in Section 3.4 as primarily a feature-engineering gap in the current survey instrument rather than a shortcoming of the random forest algorithm itself, and reinforces the mitigation already proposed there replacing coarse self-reported proxies with objective, system-logged indicators. The dataset was collected as a cross-sectional survey rather than a longitudinal evaluation, so pre-post readiness comparisons and category-transition rates were not captured. This limitation is addressed in Section 3.4, with longitudinal tracking recommended for future AI-SIM iterations.

3.4. Data and model limitations

Five interrelated limitations constrain the predictive validity and practical-impact assessment of this study; each is paired below with a concrete ICT-oriented mitigation. Sample size after deduplication: the two-level deduplication needed to prevent data leakage reduced the sample from 4,276 to 961 records, constraining statistical power. Mitigation: expand data collection across cohorts and provinces, and consider privacy-preserving federated learning across partner institutions to pool sample sizes without centralizing sensitive records.

Proxy construct validation and sector diversity: the five constructs were operationalized from available survey items rather than validated Likert-scale instruments (e.g., IT skills approximated via AI proficiency), and individually carry only modest predictive weight (0.08–0.22) (Section 3.3); sector-level diversity was also not separately encoded. Mitigation: redesign the survey around validated Likert items and

add sector as an explicit feature. Convergent evidence from regression analysis: reformulating digital readiness as a continuous index (0–100) and applying random forest regression on the same features produced $R^2 = 4.33\%$ ($MSE = 49.33$), confirming the weak signal persists regardless of classification vs. regression framing. Cross-sectional design and absence of longitudinal impact data: the student–MSME component (Section 3.3.1) lacks pre/post readiness scores or transition data, so mentoring impact is described qualitatively rather than quantified. Mitigation: future AI-SIM implementations should embed a structured pre-post assessment protocol for rigorous outcome evaluation.

System-level performance and scalability: Section 3.6's performance analysis covered only training time and inference latency on a local, non-production environment; API response time, dashboard load performance, and behavior under concurrent load were not evaluated. Mitigation: future work should add deployment-level benchmarking (e.g., Locust, Apache Bench), explore edge-computing/on-device inference for low-connectivity facilitators, and evaluate adaptive online-learning updates. These findings are reported transparently rather than adjusted to match expectations, consistent with the principle that legitimate performance bounds however modest are more valuable than inflated metrics from uncontrolled experimental design.

3.5. Implications for creative economy development

The combined findings show MSME digital transformation is shaped not only by self-reported readiness but more strongly by capital availability, legal formalization, marketplace participation, and objective scale indicators. By integrating AI analytics, geospatial intelligence, and adaptive intervention within one platform, AI-SIM supports Quintuple Helix innovation principles [27] and contributes to the discourse on digital entrepreneurial ecosystems [25], [26], independent of the classification component's current limitations.

Compared to conventional incubation, AI-SIM enables systematic readiness monitoring, personalized recommendations, and spatially informed decision support through its integrated dashboard — a scalable ICT framework for MSME digital empowerment, while transparently identifying the feature-engineering improvements needed for its predictive component. The architecture is positioned to extend toward federated learning across institutions, adaptive online-learning updates, and integration into broader regional creative-economy and entrepreneurship-education platforms.

3.6. Computational performance and system efficiency

Beyond accuracy, computational efficiency was assessed on a local workstation (Windows 10, Intel64 Family 6 Model 154 processor, 16 CPU cores) using the 768-sample training partition described in Section 2.3. Pipeline 1 required negligible training time from its single fixed-parameter fit, while Pipeline 2 incurred higher cost from the exhaustive 288-configuration hyperparameter search under 5-fold cross-validation (Table 5). Table 6 shows that AI-SIM's inference stage is lightweight relative to its training cost: a single-MSME prediction completes in 17.37 ms on average, while batch scoring of the full 193-record test set takes only 0.026 s in total (≈ 0.13 ms per sample). This latency profile is well within the responsiveness expected of a real-time incubation dashboard, confirming that the classification pipeline is practical for interactive, on-demand deployment despite the heavier one-time hyperparameter search reported in Table 5.

Table 5. Model training time

Pipeline	Configuration	Training time
Pipeline 1 (Original)	10 features, fixed hyperparameters, single fit	0.08 s
Pipeline 2 (Manuscript-Aligned)	7 features, GridSearchCV (288 candidates $\times$ 5-fold CV)	38.78 s

Table 6. Inference latency (manuscript-aligned pipeline)

Inference mode	Latency
Single prediction (1 MSME)	17.37 ms (mean over 100 predictions)
Batch prediction (193 test samples)	0.026 s total (≈ 0.13 ms per sample)

4. CONCLUSION

This study developed and evaluated AI-SIM as an integrated ICT artifact supporting MSME digital transformation while strengthening experiential entrepreneurship learning. The framework classifies MSME digital readiness into Traditional, Stable, and Digital Ready categories via random forest; after strict deduplication, the classifier achieved 33.68% and 30.57% accuracy under two configurations, both near or below the random-chance baseline. This transparent bound reflects a methodologically sound evaluation and highlights actionable directions for improving predictive performance.

Findings from 961 unique MSME records show businesses distributed relatively evenly across readiness categories, with a substantial share still Traditional. Dashboard visualizations reveal limited initial capital (mostly below IDR 50 million), incomplete legality, and uneven marketplace access. Feature importance was spread across all seven features, led by digital infrastructure and monthly revenue, with no dominant predictor, indicating that current survey instruments require refinement. Integrating ML classification, geospatial analysis, and an interactive dashboard enables targeted intervention, personalized training, and student-MSME collaboration independent of the classifier's current limitations-combining transparent, evidence-based assessment with a practical incubation ecosystem linking academia and local creative economies, positioning AI-SIM as a working smart ICT-enabled incubation platform.

Future research should prioritize: (1) redesigning survey instruments around validated Likert-scale items and explicitly encoding MSME sector; (2) expanding the sample size to 2,000–3,000 or more observations, potentially via federated learning across institutions; (3) incorporating objective, process-oriented behavior data (transaction volumes, platform usage, engagement metrics) alongside longitudinal pre-post data; and (4) advancing toward deployment-grade ICT infrastructure with load-tested APIs, edge-computing options, and adaptive online-learning. AI-SIM thus offers strong potential as a scalable, transparent digital incubation platform for MSME competitiveness and creative-economy growth, contingent on refining its predictive feature set.

ACKNOWLEDGMENTS

The authors thank the Department of Cooperatives, Small and Medium Enterprises of Lampung Province, Indonesia, especially its Head and technical staff, for enabling access to MSME data, mentoring activities, training venues, and dissemination of results — instrumental to the successful implementation of this AI-based incubation model.

FUNDING INFORMATION

This research was funded by the Ministry of Higher Education, Science, and Technology of Indonesia (Kemdiktisaintek) under the Applied Research with Model Output (Penelitian Terapan Luanan Model) scheme, Fiscal Year 2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Umar Bakti	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Maria Septijantini Alie		✓		✓		✓			✓	✓	✓			
Iin Marlina		✓	✓	✓		✓			✓		✓			

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**ding

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

INFORMED CONSENT

Informed consent was obtained from all participants; participation was voluntary and respondents were informed of the research purpose and data use for academic purposes.

ETHICAL APPROVAL

This study was conducted under institutional approval from the Department of Cooperatives, Small and Medium Enterprises of Lampung Province, following ethical standards for human-respondent data collection.

DATA AVAILABILITY

Data supporting the findings are available from the corresponding author, U.B., upon reasonable request. Due to confidentiality agreements with MSME partners and institutional regulations of the Department of Cooperatives, Small and Medium Enterprises of Lampung Province, raw datasets are not publicly accessible, as they contain sensitive business information and personal identifiers. Aggregated, anonymized, derived datasets may be provided for academic, non-commercial use, subject to institutional approval.

REFERENCES

- [1] S. Kraus, C. Palmer, N. Kailer, F. L. Kallinger, and J. Spitzer, "Digital entrepreneurship: A systematic literature review on technology-driven entrepreneurship," *Journal of Business Research*, vol. 102, pp. 11–20, 2019, doi: 10.1016/j.jbusres.2019.04.030.
- [2] M. S. S. Abad and S. Saghaei, "A cluster-based analysis of MSMEs' digital transformation barriers," *International Journal of Innovation Studies*, vol. 8, no. 1, pp. 45–58, 2024, doi: 10.1016/j.ijis.2023.11.002.
- [3] K. North, N. Aramburu, and O. J. Lorenzo, "Promoting digital transformation in SMEs," *Business Process Management Journal*, vol. 26, no. 7, pp. 1709–1732, 2020, doi: 10.1108/BPMJ-12-2019-0524.
- [4] H. Ha and C. K. P. Chuah, "Digital economy in {Southeast Asia}: challenges, opportunities and future development," *Southeast Asia: A Multidisciplinary Journal*, vol. 23, no. 1, pp. 19–35, 2023, doi: 10.1108/SEAMJ-02-2023-0023.
- [5] A. Ateeq, "Emerging economies and digital transformation: Opportunities and challenges," in *Business Sustainability with Artificial Intelligence (AI): Challenges and Opportunities*, vol. 1, E. AlDhaen, A. Braganza, A. Hamdan, and W. Chen, Eds. Cham, Switzerland: Springer Nature, 2024, pp. 129–136, doi: 10.1007/978-3-031-71526-6_12.
- [6] R. Blackburn, M. N. K. Saunders, and J. K. Loehler, "The role of business incubators in supporting MSME growth: A longitudinal study," *Small Business Economics*, vol. 58, no. 2, pp. 921–945, 2022, doi: 10.1007/s11187-021-00495-w.
- [7] A. J. McAdam, R. McAdam, and M. Dunn, "A critical review of the business incubation process in the digital age," *Technovation*, vol. 118, 2022, doi: 10.1016/j.technovation.2022.102506.
- [8] A. S. G. Gonzales and J. P. L. Tilos, "Virtual business incubation: A framework for digital startup ecosystems," in *Proceedings of the IEEE International Conference Industrial Engineering and Engineering Management*, 2023, pp. 112–117, doi: 10.1109/IEEM55224.2023.102634.
- [9] K. J. Lee and J. P. Robinson, "Scaling business incubation through automated decision support," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2025, pp. 441–450, doi: 10.1145/367890.367901.
- [10] T. R. Li, "Artificial intelligence in entrepreneurship: A review and research agenda," *Journal of Business Venturing Insights*, vol. 18, 2022, doi: 10.1016/j.jbvi.2022.e00317.
- [11] V. J. Mthunzi et al., "AI-driven customer segmentation for small businesses using machine learning algorithms," *IEEE Access*, vol. 11, pp. 44520–44535, 2023, doi: 10.1109/ACCESS.2023.3271104.
- [12] M. Razaghadeh Bidgoli, I. Raeesi Vanani, and M. Goodarzi, "Predicting the success of startups using a machine learning approach," *Journal of Innovation and Entrepreneurship*, vol. 13, no. 1, p. 80, 2024, doi: 10.1186/s13731-024-00436-x.
- [13] P. K. Senyo, K. Liu, and J. Effah, "Digital business ecosystem: Literature review and a framework for future research," *International Journal of Information Management*, vol. 47, pp. 52–64, 2019, doi: 10.1016/j.ijinfomgt.2019.01.002.
- [14] H. L. Smith, "Data analytics and AI for operational efficiency in MSMEs: A survey," *Expert Systems with Applications*, vol. 215, 2023, doi: 10.1016/j.eswa.2022.119250.
- [15] M. Tajvidi and A. Karami, "The effect of social media on firm performance of MSMEs," *Technological Forecasting and Social Change*, vol. 166, 2021, doi: 10.1016/j.techfore.2021.120626.
- [16] D. J. S. Sembiring and others, "Mapping MSME digital readiness using random forest algorithms," in *Proceedings of the International Conference on Information Management and Technology (ICIMTech)*, 2024, pp. 312–317, doi: 10.1109/ICIMTech61234.2024.00065.
- [17] W. Cao, N. T. Mai, and W. Guo, "Personalized learning and adaptive systems: {AI}-driven educational innovation and student outcome enhancement," *International Journal of Education and Humanities*, vol. 20, no. 2, pp. 173–182, 2025, doi: 10.54097/tc5k6825.
- [18] L. G. Anthopoulos and C. G. Kazantzi, "Geospatial analytics for urban MSME development," *Journal of the Knowledge Economy*, vol. 15, no. 1, pp. 240–265, 2025, doi: 10.1007/s13132-024-01890-w.
- [19] X. Zhang and R. Chen, "Geospatial clustering of MSMEs in developing regions," *Applied Geography*, vol. 145, 2022, doi: 10.1016/j.apgeog.2022.102745.
- [20] C. S. L. Tan and L. L. Low, "Random forest for classification of digital maturity," *Expert Systems*, vol. 40, no. 5, 2023, doi: 10.1111/exsy.13210.
- [21] N. S. Rathore, "Adaptive recommendation systems for business training using AI," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 45–59, 2024, doi: 10.1109/TLT.2024.3351221.
- [22] H. Lu and S. Gupta, "A dashboard for real-time MSME performance monitoring," in *Proceedings of the IEEE International Conference on Big Data*, 2022, pp. 890–895, doi: 10.1109/BigData55660.2022.10020.
- [23] Y. R. Wang, "Experiential learning in entrepreneurship education," *International Journal of Management Education*, vol. 21, 2023, doi: 10.1016/j.ijme.2023.100789.
- [24] M. S. G. Pimenta et al., "Academia-industry collaboration in digital ecosystems," *IEEE Transactions on Engineering Management*, vol. 71, pp. 1102–1115, 2024, doi: 10.1109/TEM.2023.3298710.
- [25] "{AI} and the creative economy: transforming content creation and influencer entrepreneurship," *Journal of Media Economics*, 2025, doi: 10.1080/08997764.2025.2562192.

- [26] M. Bejjani, G. Lutz, and M. Matthias, "Digital entrepreneurial ecosystems: A systematic literature review," *Technological Forecasting and Social Change*, vol. 189, p. 122372, 2023, doi: 10.1016/j.techfore.2023.122372.
- [27] E. G. Carayannis and J. J. Campbell, "The Quintuple Helix innovation model and the digital economy," *Journal of the Knowledge Economy*, vol. 12, no. 1, pp. 1–31, 2021, doi: 10.1007/s13132-021-00721-6.
- [28] G. I. Royani, N. S. Amelia, Marlina, and F. N. Cahya, "Comparative study of the C4.5 and random forest algorithms on the digitalization of MSMEs in Tegal regency," *INTI Nusa Mandiri*, vol. 20, no. 2, pp. 222–229, 2026, doi: 10.33480/inti.v20i2.7416.
- [29] A. Riansyah, M. Indriastuti, and M. A. Widiarta, "MSMEs digital transformation readiness prediction for data-driven decision support," *Journal Data Insights*, vol. 4, no. 1, pp. 141–152, 2026, doi: 10.26714/jodi.v4i1.1193.

BIOGRAPHIES OF AUTHORS



Umar Bakti, MM    is a Lecturer in Entrepreneurship at the Faculty of Business, Universitas Mitra Indonesia, Bandar Lampung, Indonesia. He holds a Bachelor's degree in Agribusiness from IPB and a Master's in Management. His research focuses on digital entrepreneurship, MSME empowerment, and AI-based startup incubation. He currently serves as Principal Investigator for an applied research project on creative economy and MSME digital transformation, funded by Indonesia's Ministry of Higher Education, Science, and Technology. In this study, he led conceptualization, methodological design, project administration, and manuscript preparation. He can be contacted at email: umarbakti@umitra.ac.id.



Dr. Maria Septijantini Alie, MM    is a Lecturer in Management at Universitas Mitra Indonesia, Bandar Lampung, Indonesia, with expertise in entrepreneurship development, MSME management, and creative economy strategy. Her research focuses on business incubation models and digital empowerment frameworks for small enterprises. She contributed to the analysis of entrepreneurial dimensions, development of incubation modules, and evaluation of the proposed AI-based incubation model. She can be contacted at email: maria_alie@umitra.ac.id.



Ilin Marlina, S.Kom, M.T.I    is a Lecturer in Information Technology at Universitas Mitra Indonesia, Bandar Lampung, Indonesia, specializing in artificial intelligence systems, web-based applications, and digital platform development. In this research, she led software development, AI system implementation, platform design, and technical validation of the AI-SIM framework, including system testing, debugging, and documentation. She can be contacted at email: marlyna@umitra.ac.id.