



Forecasting Internet adoption in Southeast Asia with machine learning: Do mobile and fixed broadband improve on temporal persistence?

Jamyang Tenzin^{1*}, Kinley Yangdon^{2*}, Dechen Wangdi³, and Thinley Choden⁴

¹⁻⁴Gedu College of Business Studies, Royal University of Bhutan, Gedu, Chhukha, Bhutan.

ORCID: Jamyang Tenzin, 0000-0002-2228-7239; Kinley Yangdon, 0009-0004-2610-2184; Dechen Wangdi, 0009-0008-4019-6534; Thinley Choden, 0009-0002-6763-3980.

Correspondence: Jamyang Tenzin, jamyangtenzin.gcbs@rub.edu.bt; Kinley Yangdon, kinleyyangdon.gcbs@rub.edu.bt

Abstract

Internet use has expanded rapidly across Southeast Asia, but national trajectories remain heterogeneous and short-horizon forecasting evidence is limited. This study evaluated whether routinely reported connectivity indicators improved one-year-ahead forecasts of the World Bank indicator “individuals using the Internet (% of population)” beyond a persistence benchmark for the 11 current ASEAN member states. A longitudinal archive contained 1,120 observed country-indicator-year records, with a 2000–2024 analytical panel. Rolling-origin evaluation covered 2016–2024, with each test year predicted only from earlier observations. The revised primary ridge-regression specification selected its penalty parameter within each outer training window by temporally ordered nested validation. Dependence-aware uncertainty used country-cluster bootstrap intervals, test-year resampling, a crossed country-by-year bootstrap sensitivity analysis, and an exact test on nine annual mean loss differentials. The nested ridge model achieved mean absolute error (MAE) 3.70 percentage points versus 5.04 for persistence, a 26.6% reduction, with RMSE 5.83, pooled R^2 0.941, and persistence-relative squared-error skill 0.343. The MAE difference was -1.34 percentage points (country-cluster 95% CI -2.13 to -0.57), although the crossed country-by-year interval included zero. On the identical 89-observation complete-case sample, adding lagged fixed broadband changed MAE from 3.852 to 3.887 (difference 0.035; country-cluster 95% CI -0.104 to 0.178), showing no clear incremental short-horizon gain. Out-of-sample permutation importance identified previous-year Internet use as the dominant predictor. These results support leakage-resistant forecasting benchmarks while cautioning against causal or AI-readiness interpretations.

Keywords: broadband; digital connectivity; Internet use; machine learning; rolling-origin forecasting; Southeast Asia

Abbreviations: ADB, Asian Development Bank; AI, artificial intelligence; ASEAN, Association of Southeast Asian Nations; ICT, information and communication technology; ITU, International Telecommunication Union; MAE, mean absolute error; OECD, Organisation for Economic Co-operation and Development; RMSE, root mean squared error; WDI, World Development Indicators.

Received: 31 August 2026. **Accepted:** 15 September 2026. **Published:** 26 September 2026 (version 1).

1. Introduction

Connectivity via Internet is an indispensable component of infrastructure for electronic commerce, digital public service delivery, telework, cloud computing, and services based on data. According to the ASEAN Digital Masterplan 2030, inclusiveness of connectivity, reliability of digital services, digital skills, and advanced technologies should be addressed together in an integrated vision for 2026-2030 (1). Consequently, the policy issue at stake is not confined to the very existence of the networks; rather, it is about expansion of Internet use in economies with diverse levels of development, geography, institutions, and history of infrastructure development.

Timor-Leste was admitted to membership in ASEAN on 26 October 2025 (2). In this respect, any current analysis of the situation in the region should take account of the fact that Timor-Leste has joined the region recently and, therefore, should be considered differently from its more experienced members regarding the quality and quantity of digital statistics.

Although there has been fast growth of connectivity in the Southeast Asian region, there are some notable differences in its infrastructure. According to the OECD data, there are large differences in broadband penetration and subscription intensity in the region, with mobile broadband dominating in terms of broadband subscriptions and fixed-broadband penetration being highly differentiated between

economies (3). Furthermore, both indicators cannot be used interchangeably with Internet usage because subscription intensity may be larger than 100 per 100 people because a person may have several subscriptions, while the Internet-use measure in the WDI is measured as the share of people who use the Internet.

The broader global context reinforces the need to distinguish infrastructure from realized use. The International Telecommunication Union estimated that 5.5 billion people, or 68% of the world population, were online in 2024, while substantial income and location gaps persisted (4). Regional work similarly treats digital inclusion as multidimensional, involving affordability, network quality, digital skills, rural access, and participation in the digital economy (6). Connectivity is also one enabling foundation for data-intensive and artificial-intelligence applications, but it is not by itself a measure of AI readiness (5).

Internet and broadband diffusion have already been examined with several aggregate forecasting and diffusion frameworks. Cross-country studies have modelled Internet diffusion using Gompertz and related S-shaped processes (21, 22), while Bass-type approaches have been applied to the global digital divide and broadband diffusion (23, 25). Broadband forecasting has also combined diffusion models with time-series methods (24), and mobile-broadband diffusion has been compared across alternative growth models (27). Within Southeast Asia, prior research has examined ICT diffusion across ASEAN (26), and ASEAN-specific studies have used Bass models to characterize and forecast Internet diffusion (28). The contribution of the present study is therefore not a claim to be the first regional Internet forecast.

The methodological question addressed here is narrower: given only information available before a forecast year, can simple public ICT indicators improve a one-year-ahead forecast of national Internet use beyond the previous year's value? A descriptive ranking or in-sample regression does not answer this question. Random train-test splitting can also leak future information into model evaluation. The present design therefore uses expanding-window rolling origins, an explicit persistence benchmark, and matched country-year comparisons.

The analysis evaluated two parsimonious feature sets. The core specification used previous-year Internet use, previous-year mobile-cellular subscriptions per 100 people, and a common time index. The augmented specification additionally used previous-year fixed-broadband subscriptions per 100 people. Ridge regression, random forest, and gradient boosting were benchmarked against persistence over rolling test years 2016–2024. Because no formal pre-registration or dated pre-specification record was available for the submitted ridge penalty, the revised primary ridge model selected its penalty parameter within each outer training set using temporally ordered nested validation; the originally submitted fixed settings are retained as reproducibility and exploratory sensitivity analyses.

This study makes three contributions. First, it provides a leakage-resistant one-year-ahead forecasting benchmark covering all 11 current ASEAN member states. Second, it makes persistence an explicit baseline, so predictive value is assessed relative to information already contained in the most recent Internet-use observation. Third, it separates predictive utility from causal interpretation and evaluates whether mobile and fixed-broadband indicators add incremental forecast information without treating predictive importance as evidence of infrastructure effects.

2. Materials and Methods

2.1 Study design, data sources, and terminology

This study used a quantitative longitudinal secondary-data design with country-year as the unit of analysis. The geographic frame comprised the 11 current ASEAN member states. The principal analytical window was 2000–2024, while earlier historical observations were retained in the raw archive for provenance and reuse. The supplied archive contained 1,120 observed country-indicator-year records: 327 observations for individuals using the Internet, 546 for mobile-cellular subscriptions, and 247 for fixed-broadband subscriptions.

The primary statistical source was the World Bank World Development Indicators (WDI). The three series were individuals using the Internet (% of population; IT.NET.USER.ZS), mobile-cellular subscriptions per 100 people (IT.CEL.SETS.P2), and fixed-broadband subscriptions per 100 people (IT.NET.BBND.P2) (11, 12). The telecommunications series are based principally on data supplied through the International Telecommunication Union. Historical fixed-broadband values in the submitted workbook were also described as having been checked against the Asian Development Bank Key Indicators Database (13). The data sources, variables, and their forecasting roles are summarized in Table 1.

The measured outcome is “individuals using the Internet (% of population).” In the title and general discussion, the term “Internet adoption” is used as shorthand for national-level uptake as represented by this reported Internet-use indicator; it is not treated as a direct measure of individual adoption behaviour. “Internet diffusion” is used when discussing the established diffusion literature, while “Internet use” is preferred when describing the empirical outcome, model target, results, and forecast errors.

2.2 Data provenance, cleaning, and lag construction

Data integrity was prioritized over cosmetic completeness. No synthetic observations, interpolation, forward filling, smoothing, or statistical imputation were used to replace unavailable source values. The 2000–2024 analytical grid

contained 275 possible country-year rows (11 countries $\times$ 25 years), of which 263 contained an Internet-use observation. After one-year lags were constructed within country, 251 rows were complete for the core feature set and 228 for the broadband-augmented feature set. Missing observations remained missing.

The reproducibility workflow rebuilds the analytical panel from the long-form raw archive rather than treating the calculated worksheet as an input. Text fields are standardized, years and values are coerced to valid numeric types, duplicate country-indicator-year records are rejected, the complete country-year grid is constructed, and the three indicators are pivoted to analytical columns. One-year lags are generated with within-country shifts, preventing values from one economy from being carried into another. The time index is defined as calendar year minus 2000.

The revised archive is designed to record row-level provenance for every observed source value, including exact source, UTC retrieval date, database release or source vintage, exact API query or raw-file identifier, SHA-256 file hash, cross-check source where applicable, reconciliation rule, and notes. The WDI retrieval script records these fields for archived API downloads. For the 247 fixed-broadband observations, however, the submitted workbook labels the source jointly as WDI/ADB and does not preserve enough metadata to identify retrospectively which individual values came from WDI, ADB, or an ADB cross-check. No unverified row-level provenance was imputed. The final archived dataset should therefore use the recovered historical ADB source/vintage if available; otherwise, a clearly identified replacement vintage should be archived and the affected analysis rerun with the source-vintage change disclosed.

2.3 Forecast design and models

The response variable was Internet use in year t . The core predictors were Internet use in year $t-1$, mobile-cellular subscriptions per 100 people in year $t-1$, and a linear time index. The broadband-augmented specification added fixed-broadband subscriptions per 100 people in year $t-1$. Country identity was excluded from the primary specification so that performance reflected transferable temporal and connectivity relationships rather than memorized country-specific intercepts.

Forecast accuracy was evaluated with an expanding-window rolling origin. For each outer test year from 2016 through 2024, the training set contained only observations from earlier years, and the fitted model then predicted all eligible countries in that test year. The origin moved forward one year and the procedure was repeated. This design mimics genuine one-step-ahead forecasting and avoids future-year leakage (14, 15).

The persistence benchmark predicted Internet use in year t as the observed value in year $t-1$. Ridge regression standardized predictors within each training fit. Because a documented pre-analysis justification for the submitted fixed

ridge penalty $\alpha=10$ was not available, the revised primary ridge analysis used temporally nested tuning. Candidate α values were 0.01, 0.1, 1, 10, 100, and 1000. Within each outer training window, candidates were evaluated on the preceding up to five temporally ordered inner validation years, with each inner fit restricted to observations earlier than its validation year. The outer test year was never used for hyperparameter selection. The selected α values were 0.01 for outer years 2016–2019 and 1 for 2020–2024.

The submitted random-forest and gradient-boosting specifications were retained as fixed exploratory benchmarks rather than retrospectively described as pre-specified. Random forest used 200 trees, maximum depth 6, minimum leaf size 2, and 80% of predictors at each split. Gradient boosting used 200 shallow trees, learning rate 0.03, maximum depth 2, minimum leaf size 2, and Huber loss. Randomized algorithms used seed 42 (16–19).

2.4 Metrics and dependence-aware inference

The primary error measure was mean absolute error (MAE), expressed in percentage points of Internet use. Root mean squared error (RMSE) was reported as a complementary measure that gives greater weight to large misses. The pooled out-of-sample coefficient of determination was defined relative to the pooled mean of the observed test outcomes and was not interpreted as skill against persistence. A separate persistence-relative squared-error statistic, $R^2_{OS,persistence}$, compared model squared error directly with the persistence benchmark. Positive values indicate lower squared error than persistence, zero indicates equal squared error, and negative values indicate worse performance.

Because forecast errors repeat within countries and within nine rolling test years, uncertainty was evaluated using dependence-aware resampling rather than treating all country-year errors as independent. The primary interval analysis used a country-cluster bootstrap that resampled entire countries with replacement and retained all test-year observations for each sampled country. Sensitivity analyses resampled complete test years and used a crossed country-by-year bootstrap with independently resampled country and year weights. Ninety-five percent confidence intervals were calculated for model MAE, persistence MAE, and their paired MAE difference. Country-year loss differentials were also averaged within each of the nine test years and assessed with an exact two-sided sign-flip test. Effect sizes and confidence intervals were treated as the principal inferential evidence.

The six originally submitted fixed-setting model-versus-persistence comparisons were treated as an exploratory family. Their exact test-year p values were adjusted with the Holm procedure. No comparison is described as formally preregistered or pre-specified in the absence of a dated protocol, analysis plan, or code record preceding inspection of the test results.

Table 1: Data sources, variables, and forecasting roles

Variable	Indicator code	Unit	Model role	Source / provenance status
Individuals using the Internet	IT.NET.USER.ZS	% of population	Target at year t ; lag at $t-1$	World Bank WDI / ITU
Mobile-cellular subscriptions	IT.CEL.SETS.P2	per 100 people	Core predictor at $t-1$	World Bank WDI / ITU
Fixed-broadband subscriptions	IT.NET.BBND.P2	per 100 people	Augmented predictor at $t-1$	WDI and historical ADB cross-check; exact row-level source/vintage must be completed before submission
Time index	Derived	years since 2000	Shared secular trend	Derived from calendar year

Abbreviations: ADB, Asian Development Bank; ITU, International Telecommunication Union; WDI, World Development Indicators. Missing source observations were retained as missing; no statistical imputation was applied. Fixed-broadband provenance must be resolved row by row in the final archived dataset.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$\text{MAE skill (\%)} = 100 \left(1 - \frac{\text{MAE}_{\text{model}}}{\text{MAE}_{\text{persistence}}} \right) \quad (3)$$

$$R^2_{\text{pooled}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_{\text{test}})^2} \quad (4)$$

$$R^2_{\text{OS,persistence}} = 1 - \frac{\sum_i (y_i - \hat{y}_{i,\text{model}})^2}{\sum_i (y_i - \hat{y}_{i,\text{persistence}})^2} \quad (5)$$

2.5 Permutation importance and robustness analyses

Permutation importance was recalculated strictly out of sample. For each outer origin, the tuned ridge model was fitted only on the historical training set, baseline MAE was calculated on that year's held-out test fold, and one predictor at a time was permuted only within the held-out test fold. Each feature permutation was repeated 1,000 times with random seed 20260914. Fold-specific importance was aggregated across the nine forecast origins, and uncertainty was summarized by resampling whole origins. Because the time index is constant within a single test year, within-fold permutation of that variable is uninformative; a zero value for that feature is therefore a property of the evaluation design rather than evidence that calendar time is substantively irrelevant.

Robustness analyses included country-level and year-level loss distributions, inspection of the largest forecast errors, exclusion of abrupt annual Internet-use changes, leave-one-country-out rolling-origin sensitivity in which the omitted country was removed from both training and testing, a separate structural-transfer leave-one-country-out exercise retained only as a distinct estimand, a simple trend-plus-persistence ordinary least-squares benchmark, clip-

ping sensitivity, and rolling empirical 90% and 95% prediction intervals. Core-versus-broadband comparisons were repeated on the identical 89 country-year complete-case observations.

Internet-use predictions were constrained to the admissible interval [0,100] after estimation. In the originally submitted 724 stored model predictions, only two predictions were exactly at a boundary because of clipping, both in the broadband-augmented ridge model; removing clipping changed that model's MAE from 4.238 to 4.258. In the revised nested-ridge primary analysis, eight of 92 raw predictions fell outside [0,100]; clipping reduced MAE from 3.802 to 3.698. Both sensitivities are reported so the practical effect of the rule is transparent.

2.6 Reproducibility environment

All revised analyses were executed in Python 3.13.5 using pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, SciPy 1.17.0, openpyxl 3.1.5, requests 2.32.5, and matplotlib 3.10.8. The reproducibility package contains pinned requirements, an environment file, runtime-version documentation, SHA-256 manifests, and a README. A concise data dictionary for all submitted workbook columns is provided in Supplementary Table S7.

The executable workflow includes scripts for WDI retrieval and archiving, cleaning and panel reconstruction, lag construction, rolling-origin fitting, prediction, clipping, dependence-aware inference, multiplicity adjustment, nested tuning, robustness analyses, permutation importance, prediction intervals, and figure generation. A one-command driver regenerates the analytical panel, model outputs, tables, figures, and file hashes. The rebuilt 275-row analytical panel matches the stored Analysis_Panel exactly, including its missing-value pattern, and all 724 submitted fixed-setting model predictions are reproduced to machine precision.

3. Results

The 2000–2024 analytical grid contained 275 possible country-year rows. Internet-use observations were available for 263 country-years; 251 observations were complete for the core lagged-feature model and 228 for the broadband-augmented model. Country coverage is sum-

marized in Table 2. Myanmar's Internet-use series in the assembled vintage ended in 2020 and Timor-Leste's ended in 2021, so the core rolling-origin test sample comprised 11 countries per year through 2020, 10 in 2021, and 9 in each of 2022–2024. Internet-use trajectories across all 11 current ASEAN member states are shown in Fig. 1.

The latest reported Internet-use values remained heterogeneous (Fig. 2). Malaysia recorded 98.0% in 2024, Brunei Darussalam 96.3%, Singapore 94.4%, Thailand 90.9%, and Viet Nam 84.2%. Indonesia, Cambodia, the Philippines, and Lao PDR were between about 65.6% and 72.8% in 2024. Myanmar's latest observation was 45.5% in 2020 and Timor-Leste's was 35.0% in 2021. Fig. 2 therefore marks these two observations with different symbols and years rather than visually presenting them as part of a same-year 2024 ranking.

The revised primary nested ridge analysis generated 92 out-of-sample country-year forecasts across nine test years. Persistence had MAE 5.039 and RMSE 7.193. Nested ridge reduced MAE to 3.698 and RMSE to 5.830, corresponding to 26.6% MAE skill. Pooled R^2 was 0.941, while persistence-relative squared-error skill was 0.343. The latter means that pooled squared error was 34.3% lower than persistence, whereas the pooled R^2 should be interpreted mainly as fit relative to the pooled test-set mean. Primary and exploratory rolling-origin performance is summarized in Table 3; pooled forecast error is illustrated in Fig. 3, and observed versus out-of-sample nested-ridge predictions are shown in Fig. 4.

The average nested-ridge MAE difference versus persistence was -1.341 percentage points. The country-cluster 95% CI was -2.132 to -0.571 and the test-year resampling interval was -2.334 to -0.231. The crossed country-by-year bootstrap interval was wider (-2.691 to 0.170), and the exact sign-flip test on the nine annual mean loss differentials gave $p=0.0664$. Thus, the average error reduction was practically meaningful, but the strength of formal inference depended on how both country and year dependence were represented. Dependence-aware interval estimates and the primary inferential results are summarized in Table 4.

Forecast skill varied substantially across years and countries. Nested ridge improved on persistence in six of nine test years but was worse in 2019, 2023, and 2024 on annual MAE; the largest improvements occurred in 2020–2022. At the country level, MAE skill was positive for nine of the 11 economies and negative for Brunei Darussalam and Singapore. This heterogeneity shows why pooled R^2 alone is insufficient for assessing short-horizon forecast performance. Annual MAE patterns are shown in Fig. 5 and Supplementary Table S1, while country-stratified forecast performance is reported in Supplementary Table S2.

For the six originally submitted fixed-setting model-versus-persistence comparisons, Holm adjustment removed conventional 0.05-level significance for every comparison. For example, the fixed core ridge exact test-year p value was 0.0156 before adjustment and 0.0938

after Holm correction. These fixed-setting comparisons are therefore reported as exploratory sensitivity results rather than as multiple independent confirmatory tests. The full multiplicity-adjusted comparison family is reported in Supplementary Table S8.

The fair comparison of fixed broadband used the same 89 complete-case test observations for both feature sets. Under nested ridge tuning, the core specification had MAE 3.852 and the broadband-augmented specification had MAE 3.887. The augmented-minus-core difference was only 0.035 percentage points, with a country-cluster 95% CI from -0.104 to 0.178. The data therefore do not show a clear incremental one-year-ahead forecasting benefit from fixed broadband after conditioning on prior Internet use, mobile subscriptions, and the time index. The matched core-versus-broadband results are summarized in Table 5.

Out-of-sample permutation importance confirmed the dominant role of previous-year Internet use. Permuting lagged Internet use within each held-out test year increased MAE by 18.774 percentage points on average, with an origin-bootstrap 95% CI of 15.344 to 21.872. The corresponding increase for lagged mobile subscriptions was 0.029 points (-0.040 to 0.097). The time index cannot be assessed by within-year permutation because it is constant within each test-year fold. The fold-level and aggregated out-of-sample permutation results are reported in Table 4 and Supplementary Table S6.

Robustness checks did not indicate that the pooled improvement was driven only by a few extreme annual changes. Excluding observations with annual Internet-use changes of at least 15 percentage points left 89 observations and increased nested-ridge MAE skill to 29.6%. When each country was removed from both rolling training and testing, MAE skill remained positive in every leave-one-country-out run, ranging from 22.4% to 32.8%. A simple trend-plus-persistence benchmark had MAE 3.883 compared with 3.698 for nested ridge. Empirical rolling prediction intervals achieved 94.6% coverage for the nominal 90% interval and 96.7% coverage for the nominal 95% interval, although these intervals were wide and are descriptive rather than guaranteed under independent-data assumptions. Abrupt-change sensitivity, rolling leave-one-country-out sensitivity, and additional benchmark, clipping, and interval checks are reported in Supplementary Tables S3b, S4, and S5, respectively.

The largest nested-ridge errors occurred during abrupt changes and reversals, including Cambodia in 2018 (absolute error 28.34 points after a 32.1-point annual increase), Cambodia in 2016 (23.44 points), Cambodia in 2019 (17.35 points after a decline), Brunei Darussalam in 2016 (16.29 points), and the Philippines in 2024 (13.84 points). These cases reinforce the difficulty of forecasting discrete revisions or sudden annual movements in a small country panel. The ten largest nested-ridge absolute errors are listed in Supplementary Table S3a.

Table 2: Country-level coverage of the 2000–2024 analytical panel

Country	First Internet year	Latest Internet year	Internet n	Mobile n	Fixed broadband n	Latest Internet %
Brunei Darussalam	2000	2024	25	25	22	96.30
Cambodia	2000	2024	25	25	23	68.48
Indonesia	2000	2024	25	25	25	72.78
Lao PDR	2000	2024	25	24	21	65.59
Malaysia	2000	2024	25	25	24	98.02
Myanmar	2001	2020	20	25	17	45.45
Philippines	2000	2024	25	25	23	67.26
Singapore	2000	2024	25	25	25	94.38
Thailand	2000	2024	25	25	23	90.87
Timor-Leste	2002	2021	18	22	21	35.04
Viet Nam	2000	2024	25	25	23	84.15

Latest Internet-use values for Myanmar (2020) and Timor-Leste (2021) are not contemporaneous with the 2024 values for most other economies.

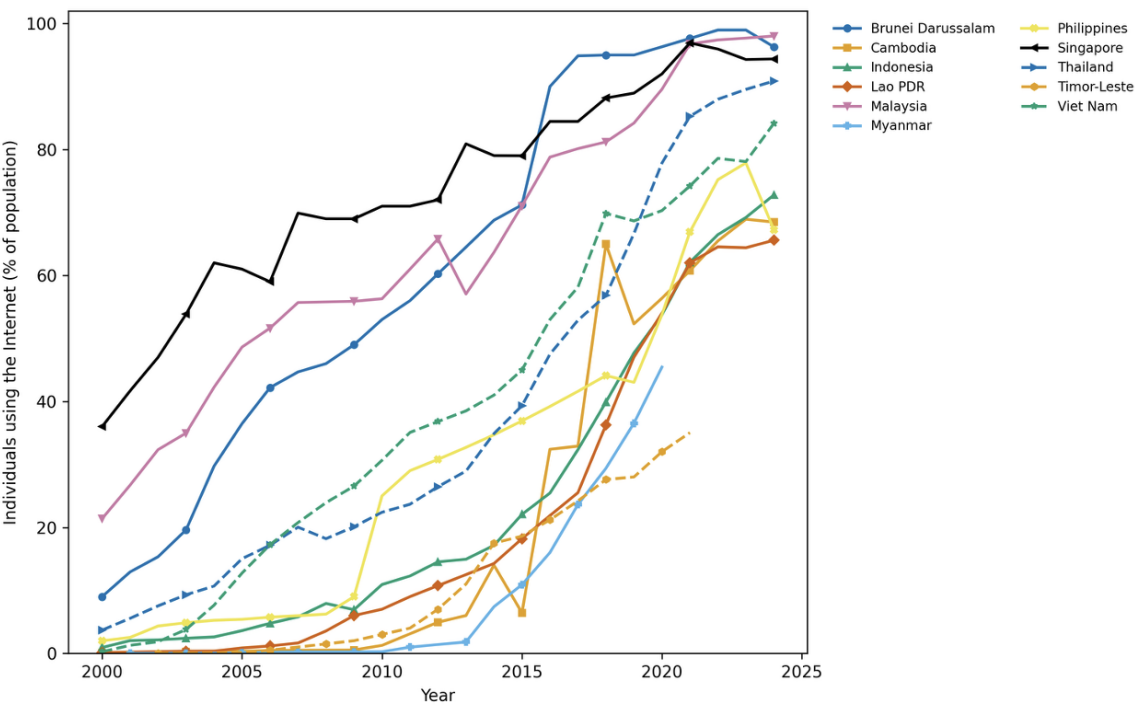


Figure 1: Internet-use trajectories across the 11 current ASEAN member states, 2000–2024. Distinct line styles and markers improve traceability; gaps and shorter series denote unavailable source years rather than imputed values. Source: World Bank WDI (11).

Table 3: Primary and exploratory rolling-origin performance for the core feature set, 2016–2024

Model	Status	n	MAE	RMSE	Pooled R ²	MAE skill %	R ² OS vs persistence
Persistence baseline	Benchmark	92	5.039	7.193	0.910	0.0	0.000
Nested ridge regression	Revised primary	92	3.698	5.830	0.941	26.6	0.343
Fixed ridge regression (α=10)	Audit sensitivity	92	3.777	5.955	0.939	25.0	0.315
Gradient boosting	Exploratory fixed setting	92	3.823	6.063	0.936	24.1	0.289
Random forest	Exploratory fixed setting	92	4.524	6.758	0.921	10.2	0.117

The nested ridge specification is the revised primary analysis because no documented pre-analysis justification for α=10 was available. Tree-model settings are retained as exploratory fixed-setting benchmarks.

4. Discussion

The practical question was whether lagged public ICT indicators improve one-year-ahead forecasts of national Inter-

net use beyond persistence. The revised results support an average improvement: temporally nested ridge regression reduced MAE from 5.04 to 3.70 percentage points. However,

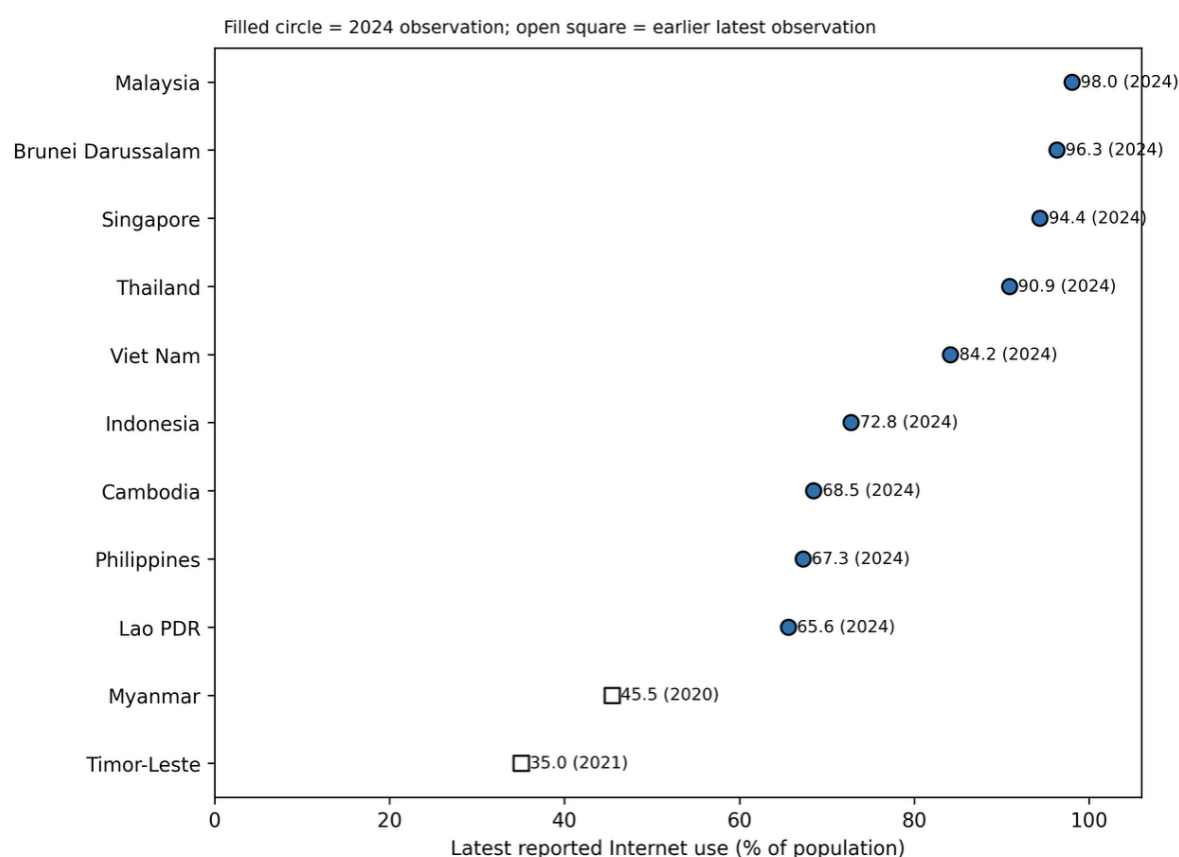


Figure 2: Latest reported Internet-use percentage by economy. Filled circles identify 2024 observations; open squares identify Myanmar (2020) and Timor-Leste (2021), preventing those earlier values from being interpreted as a same-year 2024 ranking.

Table 4: Dependence-aware inference and out-of-sample predictive importance for the revised primary ridge analysis

Item	Analysis	n	Estimate	Uncertainty / interpretation
Nested ridge vs persistence	MAE difference (model - persistence)	92	-1.341	Country-cluster 95% CI -2.132 to -0.571; year-resampling -2.334 to -0.231; crossed country $\times$ year -2.691 to 0.170; exact nine-origin $p=0.0664$
Lagged Internet use	Test-fold permutation MAE increase	92	18.774	Origin-bootstrap 95% CI 15.344 to 21.872
Lagged mobile subscriptions	Test-fold permutation MAE increase	92	0.029	Origin-bootstrap 95% CI -0.040 to 0.097
Time index	Test-fold permutation importance	92	0.000	Constant within each test-year fold; within-fold permutation is uninformative.

Permutation importance used 1,000 repeats per test fold with seed 20260914. Importance values are predictive diagnostics, not causal effects.

the dependence-aware analyses also show why the result should not be reduced to a single small p value. Country-cluster and year-resampling intervals supported a reduction, whereas the crossed country-by-year interval included zero and the exact nine-origin test was not below 0.05. Effect size, uncertainty, and stability across robustness checks are therefore more informative than the originally reported country-year Wilcoxon significance.

Temporal persistence remained the central source of predictive information. This is substantively plausible because national Internet use generally evolves gradually as infrastructure, devices, user experience, service ecosystems, affordability, and digital practices accumulate. The out-of-

sample importance analysis shows that the forecast depends strongly on previous-year Internet use, while lagged mobile subscriptions contribute much less incremental information once the recent outcome is already known. These importance values are predictive dependence diagnostics, not causal effects.

The revised fixed-broadband comparison is deliberately narrower than the submitted interpretation. The original comparison of a 92-observation core result with an 89-observation augmented result mixed different samples. On the identical 89-observation sample, nested core and augmented ridge forecasts were essentially similar, and the confidence interval for the paired MAE difference included

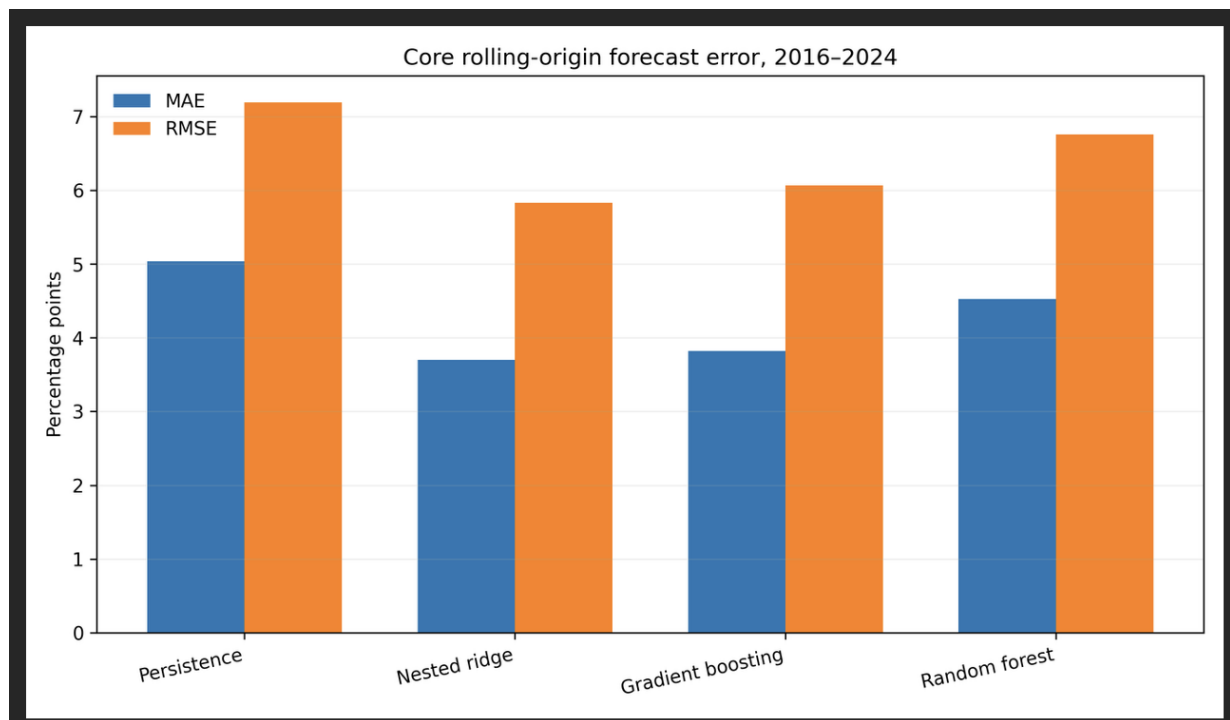


Figure 3: Pooled out-of-sample forecast error for the core feature set, 2016–2024. The nested ridge model is the revised primary specification; random forest and gradient boosting are shown as exploratory fixed-setting benchmarks. Lower MAE and RMSE indicate better accuracy.

zero. The appropriate conclusion is therefore that fixed broadband did not show a clear incremental one-year-ahead forecast benefit in this matched specification. This finding does not imply that fixed broadband is unimportant for economic development, education, cloud services, data centres, or enterprise connectivity.

The results complement, rather than contradict, regional broadband assessments. Development relevance and short-horizon incremental predictive value are different questions. A variable can be important for long-run development while contributing little additional one-year-ahead information after the previous year's outcome is already included. Similarly, mobile-subscription intensity should not be interpreted as equivalent to individual Internet use, particularly because subscriptions may exceed the population count.

The literature comparison also requires a narrower novelty claim. Internet diffusion has previously been modelled across countries with Gompertz and Bass frameworks, broadband diffusion has been forecast with diffusion and time-series methods, and ASEAN-specific work has already examined Internet diffusion using Bass models (21–28). The distinctive contribution here is the combination of all 11 current ASEAN members, explicit one-year-ahead rolling-origin evaluation, a persistence benchmark, matched feature-set comparisons, and public lagged indicators rather than the claim that Internet forecasting in the region was previously absent.

The outcome analysed here is Internet use, not AI readi-

ness. Connectivity is one enabling digital foundation for data-intensive and AI-enabled services, but AI readiness also depends on computing infrastructure, data availability and governance, skills, institutional capacity, affordability, energy reliability, sector-specific adoption, and other factors. Any relevance to AI policy is therefore contextual: persistent differences in Internet use may form part of the broader digital foundation on which AI adoption depends, but this study neither estimates nor validates an AI-readiness construct.

Country- and year-stratified results also caution against assuming uniform regional performance. The nested ridge model was worse than persistence in some countries and in several later forecast years. Large errors often coincided with abrupt increases, declines, or possible source revisions. These patterns support reporting heterogeneous loss distributions alongside pooled metrics and suggest that hierarchical or country-specific models may be valuable when longer series and additional predictors become available.

Several limitations remain. First, the analysis is predictive and observational; no coefficient, permutation score, or forecast gain identifies a causal effect of mobile or fixed-broadband infrastructure. Second, source series may be revised retrospectively, making exact source vintage and archival provenance essential. Third, mobile subscriptions measure active subscriptions rather than unique people. Fourth, national averages mask within-country differences by income, gender, age, and urban-rural location. Fifth, the

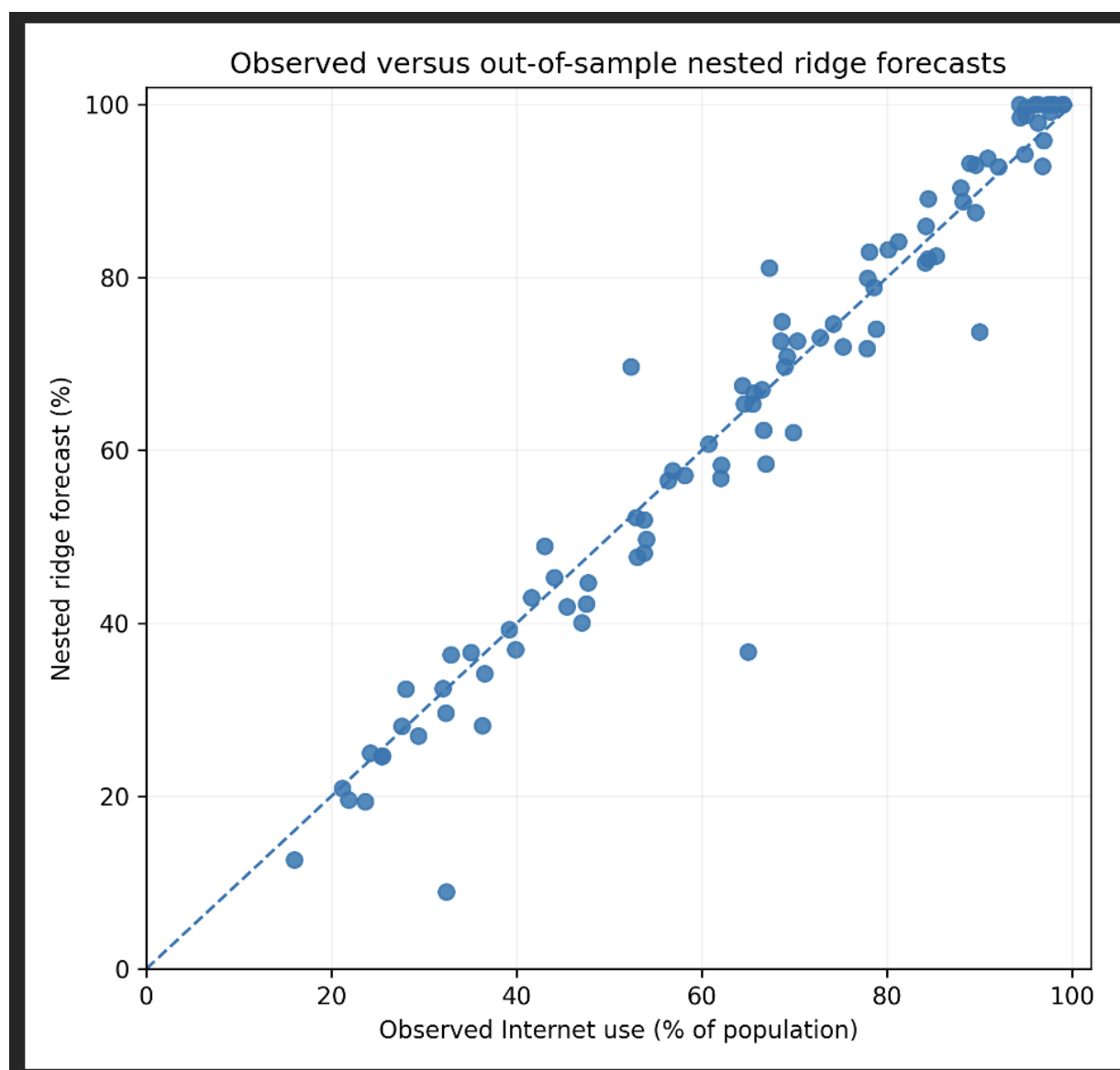


Figure 4: Observed versus out-of-sample predictions from the revised nested ridge model. Each point is one country-year rolling-origin prediction; the dashed 45-degree line represents perfect prediction.

Table 5: Matched core-versus-broadband comparison using identical test observations

Specification	n	MAE	RMSE	Pooled R^2	Direct matched comparison
Core nested ridge	89	3.852	5.986	0.935	Reference
Broadband-augmented nested ridge	89	3.887	5.993	0.935	Augmented - core MAE = 0.035; country-cluster 95% CI -0.104 to 0.178

Both specifications were trained and tested on the same 89 country-year complete-case observations.

panel is small, with only nine outer forecast origins and no more than 11 countries per origin. Sixth, data completeness differs by country and indicator, especially for fixed broadband and the later Internet-use years for Myanmar and Timor-Leste. Seventh, prediction intervals are empirical rolling intervals and should not be read as distribution-free guarantees under the panel's serial and cross-sectional dependence.

Future research could add harmonized measures of affordability, network speed, devices, education, income, competition, electricity reliability, cloud and data-centre capacity, and institutional factors when sufficiently long country-year coverage is available. Multi-step and probabilistic forecasting, hierarchical models, and subnational analyses could test whether regional pooled performance conceals important country or urban-rural differences.

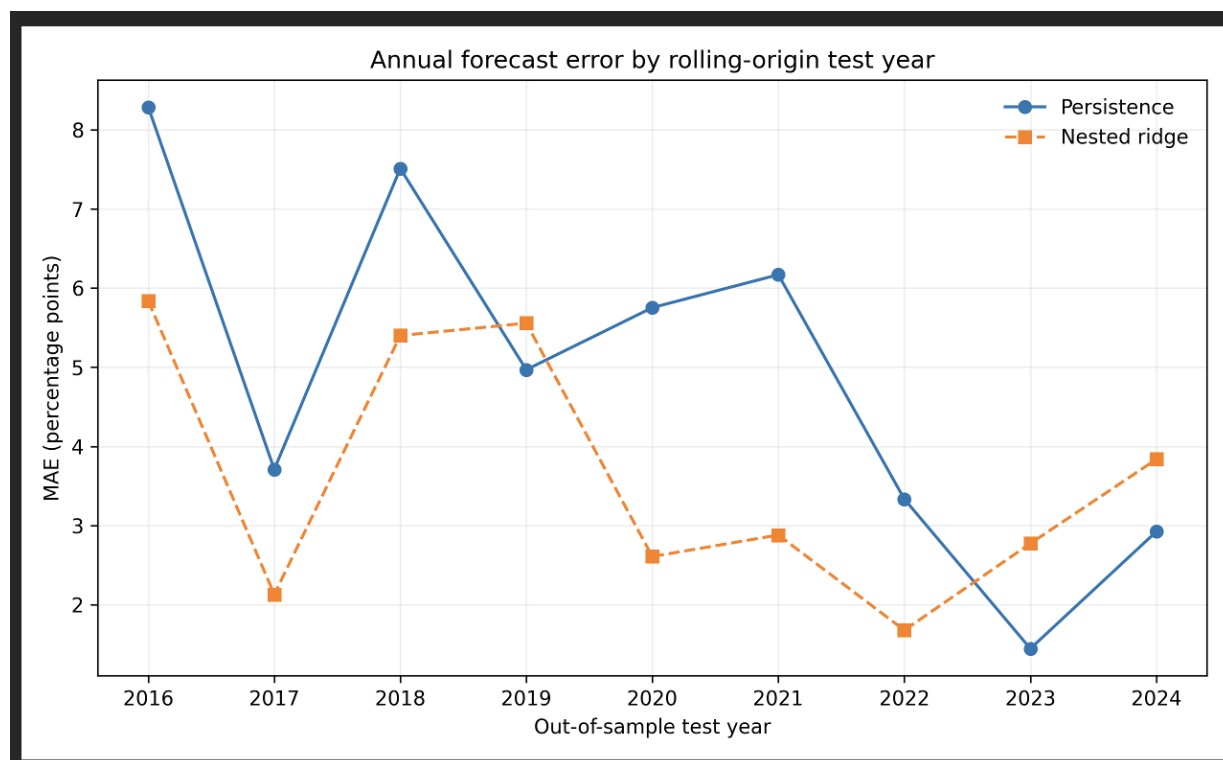


Figure 5: Annual MAE for the nested ridge model and persistence benchmark across rolling-origin test years. Ridge performs better in most, but not all, years. Test-year sample sizes decline after 2020 because of source-data incompleteness.

5. Conclusion

Using a 2000–2024 ASEAN country-year panel, this study evaluated leakage-resistant one-year-ahead forecasts of the reported percentage of individuals using the Internet. The revised primary ridge model selected its penalty through temporally ordered nested validation and reduced MAE from 5.04 percentage points for persistence to 3.70, a 26.6% average improvement. Persistence-relative squared-error skill was 0.343. Dependence-aware intervals showed that the average improvement was robust to country-cluster and year-resampling assumptions but less certain under a crossed country-by-year bootstrap, so the findings are best summarized through effect sizes, uncertainty, and robustness rather than a single significance threshold. On the identical 89-observation sample, adding fixed broadband produced no clear incremental forecasting gain. Previous-year Internet use was the dominant out-of-sample predictor. These results support disciplined forecasting benchmarks for comparative digital indicators while emphasizing three limits: predictive importance is not causality, pooled accuracy does not imply uniform country-level performance, and Internet use should not be interpreted as an AI-readiness measure.

Declarations

Authors' contributions

J.T. and K.Y. contributed equally to the conception and design of the study, data acquisition, data analysis, interpretation of the results, and drafting of the manuscript. D.W. and T.C. contributed to the critical review and revision of the manuscript. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for the work.

Acknowledgements

The authors would like to thank the World Bank and the United Nations for making the World Development Indicators and the UN E-Government Survey data publicly available. The indicators used in this study were obtained from these publicly accessible sources.

Funding

No specific funding was received for the study. This study was conducted as a part of the authors' academic and research activities.

Competing interests

The authors declare that we have no competing interests.

Ethics approval and consent to participate

Not applicable. The study used publicly available country-level secondary statistics and did not involve human participants, identifiable personal data, animals, or experimental interventions.

Consent for publication

Not applicable.

Data availability

The data, executable analysis code, analysis outputs, software environment files, and README are provided in the accompanying reproducibility package. The package includes preserved source data and code for data cleaning, lag construction, rolling-origin model fitting, prediction, clipping, dependence-aware inference, sensitivity analyses, permutation importance, and figure generation.

The final reproducibility package will be deposited in a persistent repository, and the repository name and DOI or accession number will be provided upon deposit.

Use of artificial intelligence tools

OpenAI ChatGPT was used to assist with language drafting, code generation, data-presentation planning, and document formatting. The authors retain full responsibility for the data, code, calculations, references, interpretation, and submitted text. We confirm that the authors independently verified the code, calculations, citations, statistical outputs, and final manuscript and that all authors approved the submitted version.

References

1. ASEAN. ASEAN Digital Masterplan 2030 [Internet]. Jakarta: ASEAN Secretariat; 2026 [cited 2026 Sep 14]. Available from: <https://asean.org/book/asean-digital-masterplan-2030/>
2. ASEAN. Forging a New Era: Timor-Leste Admitted into ASEAN [Internet]. Jakarta: ASEAN Secretariat; 2025 Oct 26 [cited 2026 Sep 14]. Available from: <https://tinyurl.com/fum4eb85>
3. Organisation for Economic Co-operation and Development. Extending broadband connectivity in Southeast Asia. Paris: OECD Publishing; 2023. doi:10.1787/b8920f6d-en.
4. International Telecommunication Union. Measuring digital development: Facts and Figures 2024. Geneva: ITU; 2024. Persistent link: <http://handle.itu.int/11.1002/pub/824af8b4-en>.
5. Broadband Commission for Sustainable Development. The State of Broadband 2024: Leveraging AI for Universal Connectivity. Geneva: International Telecommunication Union and UNESCO; 2024. Available from: <https://broadbandcommission.org/publication/state-of-broadband-2024/>
6. Sefrina M. An Inclusive Digital Economy in the ASEAN Region. Jakarta: Economic Research Institute for ASEAN and East Asia; 2024. Available from: <https://tinyurl.com/2ncxyp9v>
7. Suranto B, Kovač N, Haryono K, Abdul Rahman SF, Mohd Shukri AF, Suder M, et al. State of digitalization in the Southeast Asia region – bibliometric analysis. Qual Quant. 2026;60:1053–1080. doi:10.1007/s11135-025-02296-3.
8. Putra BA. Digital technologies in a post-pandemic Southeast Asia: measures for enhancing regional approaches. F1000Res. 2025;14:623. doi:10.12688/f1000research.165762.2.
9. Salman I, Ramadhan S, Arlinwibowo J, Iftitah KN, Patahuddin SM. How digital learning preparedness and ICT resources in ASEAN? Insights from PISA. J Adv Res Technol Innov Manag. 2025;15(1):17–29. doi:10.37934/jartim.15.1.1729.
10. Horn C, Rennie E. Digital access, choice and agency in remote Sarawak. Telemat Inform. 2018;35(7):1935–1948. doi:10.1016/j.tele.2018.06.006.
11. World Bank. World Development Indicators [Internet]. Washington (DC): World Bank; 2026 [cited 2026 Sep 14]. Available from: <https://databank.worldbank.org/source/world-development-indicators>
12. World Bank. About the Indicators API documentation [Internet]. Washington (DC): World Bank; [cited 2026 Sep 14]. Available from: <https://datahelpdesk.worldbank.org/knowledgebase/articles/889392>
13. Asian Development Bank. Key Indicators Database [Internet]. Manila: ADB; 2026 [cited 2026 Sep 14]. Available from: <https://kidb.adb.org/>
14. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Melbourne: OTexts; 2021. Available from: <https://otexts.com/fpp3/>
15. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data Anal. 2018;120:70–83. doi:10.1016/j.csda.2017.11.003.
16. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67. doi:10.1080/00401706.1970.10488634.

17. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32. doi:10.1023/A:1010933404324.
18. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232. doi:10.1214/aos/1013203451.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830. Available from: <https://jmlr.org/papers/v12/pedregosa11a.html>
20. Wilcoxon F. Individual comparisons by ranking methods. *Biom Bull*. 1945;1(6):80–83. doi:10.2307/3001968.
21. Kiiski S, Pohjola M. Cross-country diffusion of the Internet. *Inf Econ Policy*. 2002;14(2):297–310. doi:10.1016/S0167-6245(01)00071-3.
22. Andrés L, Cuberes D, Diouf M, Serebrisky T. The diffusion of the Internet: a cross-country analysis. *Telecommun Policy*. 2010;34(5–6):323–340. doi:10.1016/j.telpol.2010.01.003.
23. Huang CY, Chen HN. Global Digital Divide: A Dynamic Analysis Based on the Bass Model. *J Public Policy Mark*. 2010;29(2):248–264. doi:10.1509/jppm.29.2.248.
24. Christodoulos C, Michalakelis C, Varoutas D. On the combination of exponential smoothing and diffusion forecasts: an application to broadband diffusion in the OECD area. *Technol Forecast Soc Change*. 2011;78(1):163–170. doi:10.1016/j.techfore.2010.08.007.
25. Turk T, Trkman P. Bass model estimates for broadband diffusion in European countries. *Technol Forecast Soc Change*. 2012;79(1):85–96. doi:10.1016/j.techfore.2011.06.010.
26. Vu KM. ICT diffusion and production in ASEAN countries: patterns, performance, and policy directions. *Telecommun Policy*. 2017;41(10):962–977. doi:10.1016/j.telpol.2017.04.005.
27. Jha A, Saha D. Forecasting and analysing the characteristics of 3G and 4G mobile broadband diffusion in India: a comparative evaluation of Bass, Norton-Bass, Gompertz, and logistic growth models. *Technol Forecast Soc Change*. 2020;152:119885. doi:10.1016/j.techfore.2019.119885.
28. Quiban JB. Rate of internet diffusion among ASEAN countries using bass model. *Int J Appl Sci Eng*. 2021;18:2021035. doi:10.6703/IJASE.202106_18(4).002.