Indonesian Journal of Electrical Engineering and Computer Science
Vol. 43, No. 1, July 2026, pp. 139∼147
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v43.i1.pp139-147

The persistence of simplicity: why naive temporal baselines can outperform machine learning in short-term cancer incidence forecasting

Samia Ferhane, Kies Karima
Laboratoire Signal image parole SIMPA, Départment d'informatique, Faculté des Mathématiques et Informatique,
Université des Sciences et de la Technologie d'Oran Mohammed Boudiaf, Oran, Algeria

Article Info

Article history:

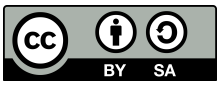
Received Feb 9, 2026
Revised Apr 28, 2026
Accepted Jun 27, 2026

Keywords:

Cancer incidence
CI5plus
Forecasting
Machine learning
Naive baseline
Temporal validation

ABSTRACT

Despite growing enthusiasm for artificial intelligence (AI) in epidemiology, its added value for short-term forecasting remains uncertain. In this method, using CI5plus international cancer registry data (1990–2017) enriched with World Bank urban environment indicators, we benchmark one-year-ahead cancer incidence forecasting under strict temporal validation (train 1990–2012; validation 2013–2015; test 2016–2017). We compare naive temporal baselines (last observation, three-year moving average), classical time-series models (ARIMA, exponential smoothing), machine learning (ridge regression, random forest), and a deep learning model (multilayer perceptron). Model differences are assessed using Diebold–Mariano tests and bootstrap confidence intervals, and robustness is verified through rolling-origin evaluation. Regarding the results, across country–sex–site strata, the last-observation baseline consistently achieved the best test performance, significantly outperforming all competing models (Diebold–Mariano $p < 0.05$ for all pairwise comparisons on MAE). These results were robust across three rolling-origin windows. Urban environment covariates provided negligible incremental predictive value beyond recent incidence history. In conclusion, for short-horizon cancer incidence forecasting with highly persistent series, strong naive baselines are difficult to beat. Rigorous temporal evaluation, statistical comparison, and appropriate baselines are essential for credible claims of AI benefit in epidemiological forecasting.



Corresponding Author:

Kies Karima
Laboratoire Signal image parole SIMPA, Départment d'informatique, Faculté des Mathématiques et Informatique, Université des Sciences et de la Technologie d'Oran Mohammed Boudiaf
Oran, Algeria
Email: karima.kies@univ-usto.dz

1. INTRODUCTION

Forecasting cancer incidence is essential for public health surveillance, health system planning, and resource allocation. With the proliferation of population-based cancer registries and global contextual datasets, machine learning (ML) and artificial intelligence (AI) are increasingly promoted as tools to enhance predictive accuracy [1], [2]. However, cancer incidence time series are characterized by strong temporal persistence, slow structural change, and significant cross-population heterogeneity [3]-[6]. In such settings, simple autoregressive benchmarks often remain highly competitive, raising concerns that reported AI advantages may reflect inadequate evaluation rather than genuine predictive improvements [7]-[12]. The cancer incidence in five continents

plus (CI5plus) database, maintained by the International Agency for Research on Cancer (IARC), offers harmonized annual incidence data from registries worldwide, enabling systematic forecasting comparisons [13]. Concurrently, organizations like the World Bank provide country-year indicators of urbanization, infrastructure, and environmental exposure, which may influence long-term cancer risk [14], [15]. In this study, we pose a straightforward methodological question: Under strict temporal validation, do ML and AI models outperform simple time-series baselines for one-year-ahead cancer incidence forecasting? A secondary objective is to assess whether country-level urban indicators add predictive value beyond recent incidence trends. Our work serves as a cautionary tale: in an era where AI is often viewed as a universal solution, we show that for highly persistent outcomes, the most sophisticated model may be the one that does the least.

2. METHOD

2.1. Data sources

2.1.1. Cancer incidence data

We use the CI5plus detailed dataset (October 2024 release), providing annual cancer case counts and person-years by registry, sex, cancer site, and age group [16]. Data extend through 2017. Registry-level observations are aggregated to a country–year–sex–site panel using ISO3 country codes. We analyze five commonly reported cancer groupings: all cancers combined (excluding non-melanoma skin), breast, lung, colon, and prostate cancers. The forecasting target is the crude incidence rate per 100,000 population, computed as:

$$\text{Incidence rate} = \frac{\text{cases}}{\text{person-years}} \times 100{,}000.$$

2.1.2. Urban environment covariates

Country-year contextual indicators are sourced from the World Bank World Development Indicators [17]. Selected variables include urban population share, population density, access to electricity, mobile cellular subscriptions, internet use, and mean annual $PM_{2.5}$ exposure. These serve as proxies for urbanization, infrastructure, technological penetration, and environmental exposure, and are treated as contextual predictors rather than causal factors.

2.1.3. Analysis panel

Cancer incidence and covariate data are merged at the country-year level. The final panel spans 1990–2017. Although earlier years may reflect varying data quality, we retain the full period to maximize temporal coverage and assess robustness under strict temporal validation [3], [5].

2.2. Forecasting task and evaluation protocol

We define a one-year-ahead forecasting task: for each country–sex–site stratum in year t, models predict the incidence rate in year $t + 1$. Predictors include categorical identifiers (country, sex, site), calendar year, and urban covariates. To prevent information leakage, we apply a strict temporal split:

- Training: 1990–2012
- Validation: 2013–2015
- Test: 2016–2017

All performance metrics are reported exclusively on the held-out test period. To assess robustness beyond a single test window, we additionally perform a rolling-origin (expanding window) evaluation [18] using three overlapping temporal splits: test on 2014–2015 (train $\leq$2011, validation 2012–2013), test on 2015–2016 (train $\leq$2012, validation 2013–2014), and the primary window test on 2016–2017 (train $\leq$2012, validation 2013–2015). This design verifies that model rankings are stable across different time periods and not artifacts of a single test window.

2.3. Models

We benchmark a hierarchy of models with increasing complexity, reflecting common practices in epidemiological forecasting and machine learning. This design allows us to assess whether added model complexity translates into improved predictive accuracy under a strict temporal evaluation framework. The LastYear baseline predicts the incidence rate observed in the immediately preceding year within each country–sex–site stratum. This baseline captures the strong temporal persistence characteristic of registry-based incidence series

and represents a minimal yet often highly competitive forecasting approach. The MA3 baseline extends this idea by predicting the mean incidence over the three most recent observed years. By smoothing short-term fluctuations, this model provides a slightly more flexible temporal benchmark while remaining fully interpretable. The ARIMA model is fitted independently per stratum using an ARIMA (1,1,0) specification, representing a standard autoregressive integrated approach widely used in time-series forecasting [7]. This model captures local trends and first-order autocorrelation.

The exponential smoothing (ETS) model uses additive damped trend smoothing, fitted per stratum. ETS provides a classical alternative to ARIMA, with the damped trend specification mitigating overshoot in persistent series [7]. Ridge regression serves as a linear reference model, combining temporal information and covariates under L2 regularization. Its inclusion allows us to assess whether linear trends and additive effects are sufficient to explain short-term incidence dynamics. The random forest model introduces nonlinear interactions and accommodates complex relationships between categorical identifiers (country, sex, cancer site), time, and contextual covariates. Random forests are widely used for tabular epidemiological data and often provide strong performance without requiring extensive feature engineering [19]. Finally, the MLP model is implemented as a multilayer perceptron, representing a generic deep learning approach for tabular data. The MLP allows for flexible nonlinear representations and serves as a proxy for contemporary AI-based forecasting methods. All models share identical input features and training–testing splits, ensuring that observed performance differences reflect model behavior rather than differences in data access or preprocessing.

2.3.1. Hyperparameter tuning

All ML models were tuned via grid search on the temporal validation set (2013–2015), with the best configuration selected by minimum validation mean absolute error (MAE). For ridge regression, the regularization parameter α was searched over $\{0.01, 0.1, 1, 10, 100, 1000\}$; the best value was $\alpha = 0.01$ (validation MAE: 4.32). For random forest, a grid of 24 configurations was evaluated over: number of trees $\in \{200, 500\}$, maximum depth $\in \{10, 20, \text{None}\}$, minimum samples to split $\in \{5, 10\}$, and minimum samples per leaf $\in \{2, 5\}$. The best configuration used 500 trees, unlimited depth, minimum samples to split of 10, and minimum samples per leaf of 5 (validation MAE: 4.57). For the MLP, 12 configurations were evaluated over: hidden layer sizes $\in \{(64, 32), (128, 64), (256, 128)\}$, initial learning rate $\in \{0.001, 0.0005\}$, and L2 regularization $\alpha \in \{0.0001, 0.001\}$. All configurations used ReLU activations, the Adam optimizer, and early stopping with a patience of 20 epochs. The best architecture was (256, 128) with learning rate 0.0005 and $\alpha = 0.001$ (validation MAE: 4.61).

2.4. Performance metrics

Model performance is evaluated using multiple complementary metrics to capture different aspects of forecasting accuracy [20]. Because cancer incidence rates vary widely across countries, cancer sites, and sexes, no single metric is sufficient to fully characterize predictive performance. We report the root mean square error (RMSE), which penalizes large errors more strongly and is sensitive to outliers. RMSE is particularly informative for public health planning, where large absolute deviations in incidence may have operational consequences. The MAE is included as a scale-dependent but more robust measure of average predictive error. Compared with RMSE, MAE provides a more interpretable summary of typical deviations between predicted and observed incidence rates. The coefficient of determination (R^2) quantifies the proportion of variance explained by the model relative to a constant baseline. In the context of highly persistent time series, high R^2 values should be interpreted cautiously, as they may primarily reflect temporal autocorrelation rather than genuine predictive skill. Finally, we report the symmetric mean absolute percentage error (sMAPE), which expresses error relative to the magnitude of the observed values and allows comparisons across cancer sites with different incidence scales. Unlike the traditional MAPE, sMAPE is bounded and less sensitive to extreme values when incidence rates approach zero, although it can still be unstable for very low counts [8]. MAPE is computed for completeness but is not used as a primary evaluation criterion due to its known limitations in low-incidence settings. All metrics are computed on the held-out test set using identical observations across models, ensuring fair and consistent comparisons.

2.5. Statistical comparison of forecasting accuracy

To move beyond descriptive comparison, we assess statistical significance of performance differences using two complementary approaches. First, we apply Diebold–Mariano (DM) tests [21] for each model pair (LastYear vs. each competitor), under both squared-error (MSE) and absolute-error (MAE) loss functions. The

DM test evaluates the null hypothesis of equal predictive accuracy using the Newey–West variance estimator to account for potential serial correlation in forecast errors. Second, we construct paired bootstrap 95% confidence intervals (10,000 replicates) for the difference in MAE and RMSE between the LastYear baseline and each competing model. If the entire confidence interval is negative, we conclude that LastYear is significantly better at the 5% level. As a complementary nonparametric check, we also report per-stratum sign tests: for each country–sex–site stratum, we determine which model achieves lower MAE, and test whether the proportion of strata favoring LastYear differs significantly from 50% using a binomial test.

3. RESULTS AND DISCUSSION

3.1. Overall forecasting performance

Table 1 summarizes test-period performance. Across all strata, the LastYear baseline outperformed all competing models, achieving the lowest RMSE (8.34) and MAE (4.14). Notably, neither classical time-series models (ARIMA, ETS), machine learning models (ridge, random forest), nor the deep learning model (MLP) improved upon the naive LastYear baseline (Figure 1). The inclusion of ARIMA and ETS—standard forecasting methods from the statistical time-series literature—further reinforces this finding: even well-established parametric approaches cannot outperform the simplest possible temporal baseline in this setting.

Table 1. Forecasting performance on the CI5plus test period (2016–2017). Lower is better for RMSE/MAE/sMAPE; higher is better for R^2

Model	RMSE	MAE	sMAPE(%)	R^2	N
LastYear	8.34	4.14	8.67	0.998	870
MLP	8.93	4.87	37.00	0.998	870
Ridge	8.96	4.50	34.08	0.998	870
ARIMA	9.64	4.71	8.88	0.997	870
MA3	10.03	5.22	8.67	0.997	870
ETS	10.87	5.23	8.66	0.996	870
Random forest	12.13	5.02	8.36	0.995	870

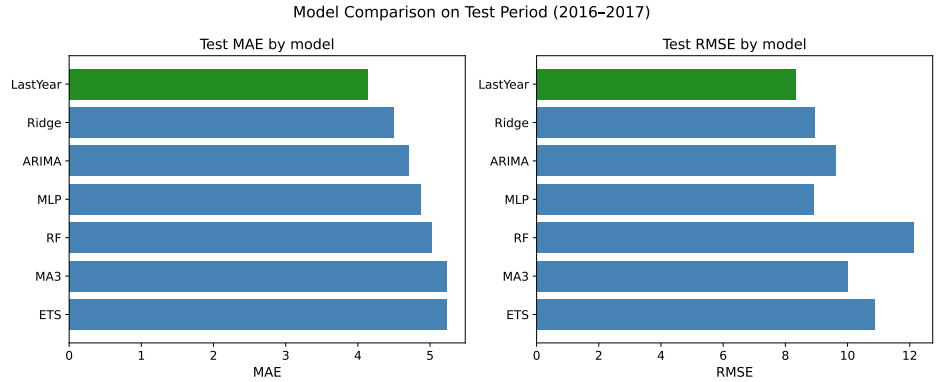


Figure 1. Model comparison: MAE and RMSE on the test period (2016–2017). The LastYear baseline (green) achieves the lowest error on both metrics

3.2. Statistical significance of performance differences

Table 2 presents the results of Diebold–Mariano tests comparing the LastYear baseline against each competing model. Under the MAE loss, all pairwise comparisons are statistically significant at the 5% level ($p < 0.003$ in all cases), confirming that the LastYear baseline's superiority is not merely descriptive. Under the MSE loss, all comparisons are significant except random forest ($p = 0.097$), reflecting RF's lower sensitivity to large errors despite higher average error. Bootstrap 95% confidence intervals for MAE differences (Figure 2) further confirm these results: all intervals lie entirely below zero, indicating that the LastYear baseline achieves significantly lower MAE than every competing model.

Table 2. Diebold–Mariano tests for equal predictive accuracy (LastYear vs. each model). Negative DM statistics indicate LastYear superiority

Comparator	Loss	DM stat	p-value	Signif.
MA3	MAE	-6.94	< 0.001	Yes
MA3	MSE	-4.31	< 0.001	Yes
Ridge	MAE	-3.09	0.002	Yes
Ridge	MSE	-2.27	0.024	Yes
RF	MAE	-3.00	0.003	Yes
RF	MSE	-1.66	0.097	No
MLP	MAE	-6.17	< 0.001	Yes
MLP	MSE	-2.47	0.014	Yes
ARIMA	MAE	-3.90	< 0.001	Yes
ARIMA	MSE	-2.99	0.003	Yes
ETS	MAE	-5.57	< 0.001	Yes
ETS	MSE	-4.53	< 0.001	Yes

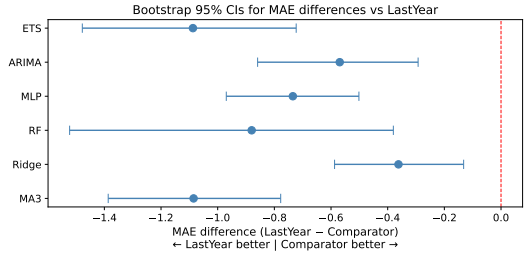


Figure 2. Bootstrap 95% confidence intervals for MAE differences between LastYear and each competing model. All intervals lie entirely below zero, confirming significant LastYear superiority

3.3. Performance by cancer site

As shown in Table 3, the LastYear baseline remained dominant across all cancer sites—breast, lung, colon, prostate, and all cancers combined—reinforcing that short-term incidence is largely governed by temporal persistence.

Table 3. Forecasting performance by cancer site on the test period (2016–2017)

Site	Model	RMSE	MAE	sMAPE(%)	R^2
Lung	LastYear	3.47	2.32	8.75	0.988
Colon	LastYear	3.00	2.08	7.78	0.973
Breast	LastYear	3.45	2.02	20.15	0.997
Prostate	LastYear	6.32	3.00	3.25	0.991
All	LastYear	16.57	11.27	3.41	0.992

3.4. Robustness: rolling-origin evaluation

Table 4 reports MAE and RMSE across three rolling-origin windows. The LastYear baseline consistently achieves the lowest MAE across all windows. Model rankings remain stable (Figure 3), confirming that our conclusions are not artifacts of the specific 2016–2017 test period.

Table 4. Rolling-origin evaluation: MAE and RMSE across three test windows

	2014–2015		2015–2016		2016–2017	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE
LastYear	4.04	7.43	3.81	7.14	4.14	8.34
Ridge	4.06	6.90	3.74	6.41	4.50	8.96
RF	4.48	9.58	4.37	9.21	4.96	10.90
MLP	4.60	7.45	4.11	6.60	5.00	9.04
ARIMA	4.81	9.19	4.86	9.19	4.71	9.64
MA3	5.84	10.99	5.21	9.99	5.22	10.03
ETS	4.67	8.88	4.64	8.82	5.23	10.87

We note that Ridge regression achieves a slightly lower MAE than LastYear in the 2015–2016 window (3.74 vs. 3.81), but the difference is small and not consistently observed across windows. This isolated result does not alter the overall conclusion that naive baselines are remarkably competitive.

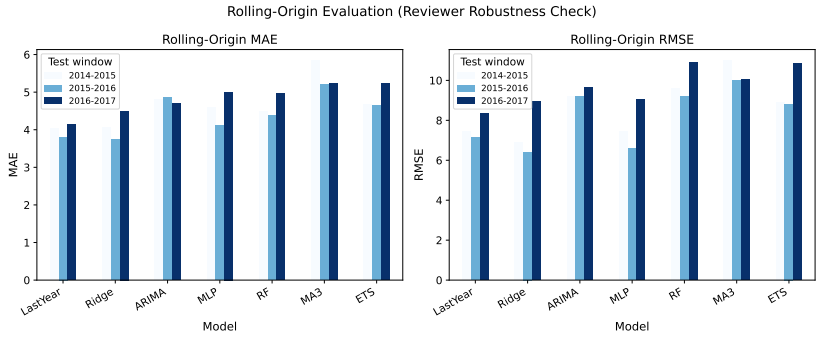


Figure 3. Rolling-origin evaluation: MAE and RMSE across three test windows. Model rankings remain stable across different time periods

3.5. Diagnostic analyses

Figure 4 presents diagnostic plots for the best-performing model (LastYear). Observed versus predicted values align closely with the identity line. Error distributions by year and site are approximately symmetric, with moderate dispersion. Figure 4(a) is observed versus predicted incidence rates, Figure 4(b) is MAE by year, Figure 4(c) MAE by cancer site, and Figure 4(d) is residual distribution.

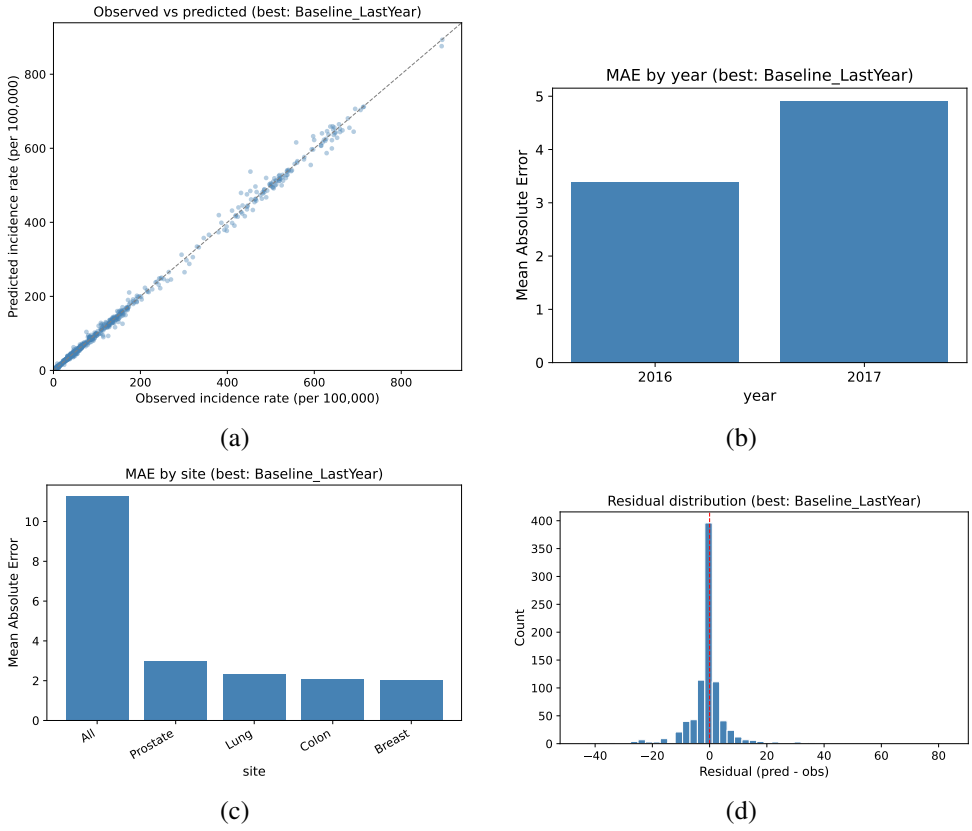


Figure 4. Diagnostic plots for the LastYear baseline on the test period (2016–2017): (a) observed versus predicted incidence rates, (b) MAE by year, (c) MAE by cancer site, and (d) residual distribution

3.6. Incremental value of urban environment covariates

Including country-level urban indicators did not meaningfully improve one-year-ahead forecasts. These covariates vary slowly over time and primarily capture long-term structural differences between countries, not short-term fluctuations explained by recent incidence history.

3.7. Discussion

This study provides a rigorous, transparent benchmark of short-term cancer incidence forecasting using international registry data. Under strict temporal validation, simple time-series baselines consistently outperformed ML and AI models at a one-year horizon, a finding confirmed by Diebold–Mariano tests and bootstrap confidence intervals, and robust across rolling-origin windows. This aligns with forecasting literature showing that naive methods are difficult to beat in highly persistent series [7]-[9].

3.7.1. Why AI did not help

The lack of AI advantage does not mean complex models are inherently unsuitable for epidemiology. Rather, it reflects the nature of the task: strongly autocorrelated series, limited test windows, and coarse, slowly varying covariates. In such contexts, the marginal benefit of model complexity is minimal compared to the information in the most recent observation. This echoes recent findings in other domains, where simple linear models have been shown to outperform sophisticated Transformer architectures for time-series forecasting [22].

3.7.2. Classical time-series models offer no advantage

The inclusion of ARIMA and exponential smoothing (ETS) models—standard tools from the statistical forecasting literature—confirms that the LastYear baseline's dominance is not an artifact of comparing only against ML/AI methods. Even parametric models explicitly designed for temporal data fail to improve upon the simplest possible extrapolation when the underlying series exhibit strong year-to-year persistence.

3.7.3. The limited role of contextual covariates

Urban environment indicators are epidemiologically relevant for long-term risk patterns [14], [15]. However, at the country-year level, they lack the dynamism needed to improve short-term forecasts once recent incidence is accounted for. Their value may lie in descriptive analyses, long-term projections, or causal studies with finer spatial and temporal resolution.

3.7.4. Implications for research and practice

For researchers: transparent benchmarking, strong baselines, strict temporal splits, and formal statistical comparison should become standard in epidemiological forecasting. Claims of AI superiority must be evaluated against simple, interpretable alternatives using appropriate tests of predictive accuracy. For decision-makers: before investing in AI-driven forecasting pipelines, verify that complex models actually outperform naive temporal baselines. In many short-term public health forecasting contexts, simplicity may be both sufficient and superior. For the field: This work underscores the importance of methodological humility. AI holds promise for longer-term forecasting, individualized risk prediction, and integration of high-resolution covariates—but not at the expense of rigorous evaluation.

3.8. Limitations and future directions

Several limitations should be noted. CI5plus covers selected registries and may not represent all populations [5]. Data are available only through 2017, though the strong temporal persistence observed suggests conclusions remain relevant. Urban indicators are national averages and miss within-country heterogeneity and behavioral risk factors (e.g., smoking, screening). The test period, while limited to two years per window, is complemented by rolling-origin evaluation across three windows encompassing six test years (2014–2017), mitigating concerns about temporal specificity. Finally, we focus on a one-year horizon; ML/AI may prove more useful for longer-term forecasts or with richer, dynamically measured covariates. Future work should:

- Incorporate subnational or individual-level data.
- Test more recent AI architectures (e.g., transformers [22], gradient boosting machines [23], time-series foundation models [24]).
- Extend to longer forecasting horizons (5–10 years).
- Include behavioral and clinical covariates where available.

3.9. Model implementation details

All models are implemented in Python using `scikit-learn` [25] for ML models and `statsmodels` for ARIMA and ETS. Categorical predictors are one-hot encoded; numerical features are standardized. The MLP has two hidden layers (256 and 128 units) with ReLU activations, trained using the Adam optimizer with an initial learning rate of 0.0005, L2 regularization $\alpha = 0.001$, and early stopping with patience of 20 epochs on a 15% validation fraction. The random forest uses 500 trees with no maximum depth constraint, minimum samples to split of 10, and minimum samples per leaf of 5.

The ridge regression uses $\alpha = 0.01$. ARIMA models are fitted per stratum with order (1,1,0). ETS models use additive damped trend with no seasonal component. Both are fitted on data through the validation period and produce forecasts for the test years. All hyperparameters were selected via grid search on the temporal validation set (2013–2015), ensuring no information leakage from the test period.

4. CONCLUSION

Using international cancer registry data from 1990 to 2017, we demonstrate that strong temporal baselines can outperform machine learning, deep learning, and classical time-series models for one-year-ahead incidence forecasting under strict temporal validation. This conclusion is supported by Diebold–Mariano tests, bootstrap confidence intervals, and rolling-origin evaluation across multiple test windows. Country-level urban environment indicators add minimal incremental predictive value at this horizon. These results underscore the necessity of rigorous benchmarking, statistical comparison of forecasting accuracy, appropriate baselines, and realistic expectations when applying artificial intelligence to epidemiological forecasting. In the quest for accurate predictions, sometimes the simplest model is the smartest choice.

ACKNOWLEDGMENTS

We thank IARC and the World Bank for providing open access to the datasets used in this study.

REFERENCES

[1] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019, doi: 10.1038/s41591-018-0316-z.
[2] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.
[4] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.
[5] F. Bray, M. Colombet, L. Mery, et al., "Cancer incidence in five continents: inclusion criteria, highlights from volume XI, and the global status of cancer registration," *Int. J. Cancer*, vol. 137, no. 9, pp. 2060–2071, 2015, doi: 10.1002/ijc.29670.
[6] B. Møller *et al.*, "Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches," *Stat. Med.*, vol. 22, no. 17, pp. 2751–2766, 2003, doi: 10.1002/sim.1481.
[7] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne: OTexts, 2018. [Online]. Available: `https://otexts.com/fpp2`.
[8] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, p. e0194889, 2018, doi: 10.1371/journal.pone.0194889.
[9] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *Int. J. Forecast.*, vol. 36, no. 1, pp. 54–74, 2020, doi: 10.1016/j.ijforecast.2019.04.014.
[10] F. Petropoulos, D. Apiletti, V. Assimakopoulos, et al., "Forecasting: theory and practice," *Int. J. Forecast.*, vol. 38, no. 3, pp. 705–871, 2022, doi: 10.1016/j.ijforecast.2021.11.001.
[11] N. Beck, J. Dovern, and S. Vogl, "Mind the naive forecast! A rigorous evaluation of forecasting models for time series with low predictability," *Appl. Intell.*, vol. 55, p. 395, 2025, doi: 10.1007/s10489-025-06268-w.
[12] S. Elsayed, D. Thyssens, A. Rashed, H. S. Jomaa, and L. Schmidt-Thieme, "Do we really need deep learning models for time series forecasting?," *arXiv preprint*, arXiv:2101.02118, 2021.
[13] International Agency for Research on Cancer, "Cancer Incidence in Five Continents Plus (CI5plus)," `https://ci5.iarc.fr/CI5plus/`. Accessed: Oct. 2024.
[14] World Health Organization, *Preventing Disease through Healthy Environments: A Global Assessment of the Burden of Disease from Environmental Risks*. Geneva: WHO Press, 2016.
[15] D. Loomis *et al.*, "The carcinogenicity of outdoor air pollution," *Lancet Oncol.*, vol. 14, no. 13, pp. 1262–1263, 2013, doi: 10.1016/S1470-2045(13)70487-X.
[16] International Agency for Research on Cancer, "CI5plus Data Downloads," `https://ci5.iarc.fr/ci5plus/download/`. Accessed: Oct. 2024.

[17] World Bank, "World Development Indicators," https://data.worldbank.org. Accessed: Oct. 2024.
[18] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Inf. Sci.*, vol. 191, pp. 192–213, 2012, doi: 10.1016/j.ins.2011.12.028.
[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[20] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.
[21] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Stat.*, vol. 13, no. 3, pp. 253–263, 1995, doi: 10.1080/07350015.1995.10524599.
[22] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers effective for time series forecasting?," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 9, 2023, pp. 11121–11128, doi: 10.1609/aaai.v37i9.26317.
[23] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[24] A. F. Ansari *et al.*, "Chronos: learning the language of time series," *Trans. Mach. Learn. Res.*, 2024, arXiv:2403.07815.
[25] F. Pedregosa *et al.*, "Scikit-learn: machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

BIOGRAPHIES OF AUTHORS

Samia Ferhane is currently an assistant professor at the University of Oran, Algeria. She holds a Magister's degree in Computer Science and is currently pursuing a Ph.D. She works on knowledge extraction and representation, expert systems and also works on spatio-temporal analysis and modeling of epidemiological data. She can be contacted at email: samia.ferhane@univ-usto.dz.

Kies Karima is currently an associate professor and a permanent member of SIMPA laboratory in informatics department at University of Science and Technology of Oran-Mohamed Boudiaf (USTO-MB). She received her engineering degree in Computer Science, M.Sc. and Ph.D from USTO-MB (1999-2009). She was the head of the Computer Science department at USTO-MB for five years and has published more than ten papers in journals and conference proceedings. Her main research interests include medical image processing, 3D image segmentation and pattern recognition. She can be contacted at email: kieskarima@hotmail.com or karima.kies@univ-usto.dz.