TRIZ-Informed Validation of Predictive AI in a Holistic Cybersecurity Sandbox: A Machine Learning Pipeline for Trustworthy Forecasting

Isabella Fodor[1], Andreea Ionica[1], Monica Leba[1], Ionela Samuil[1], Eugen Florian Popescu[2], Nicusor Gelu Constantin[2], Dragos Dabija[3]
[1]University of Petrosani, Universitatii 20, Petrosani, 332006
[2]Aria United S.R.L., str. Theodor Sperantia nr 104, Bucuresti, 030939
[3]SOFTTEHNICA SRL, str. Splaiul Unirii 76, Bucuresti, 040037
Romania

Received: March 14, 2026. Revised: June 16, 2026. Accepted: July 16, 2026. Available online: November 18, 2026.

Abstract—We propose a TRIZ-Informed Machine Learning (TIML) framework as part of an integrated cybersecurity sandbox to examine and validate AI-driven decision support systems. The framework employs six of the TRIZ creative principles as a sophisticated AI vulnerability diagnostic protocol. This framework investigates and addresses several specific adversarial concerns: COVID-19-induced structural breaks and cross-continuous entity scale heterogeneity. The sandbox validation approach improves random forest validation R^2 from 0.7251 to 0.8978, and MAPE improves from 11.99% to 6.83%. This work develops a method for a standards-oriented validation and evaluation framework within the bounds of the EU's AI Act and combines the principles of causal irreversibility and temporal autocorrelation. The TIML framework is introduced as a modular sandbox testing framework for use with other predictive AI systems that are operating in adversarial conditions and are seeking real-world validation.

Keywords— trustworthy AI validation, TRIZ inventive principles, ensemble machine learning, tourism demand forecasting, TRIZ-informed, machine learning.

I. INTRODUCTION

THE engineering of trustworthy, robust, and auditable artificial intelligence systems in advanced and sophisticated real-world operation settings is one of the most difficult problems facing AI engineering today. The increasing deployment of AI systems in real-world operational settings of increased complexity and uncertainty poses problems in designing and validating the systems before deployment. A unified evaluation architecture integrates technical assessment of systems, adversarial stress testing, simulation-based validation, quality management, and compliance verification, as well as a human-in-the-loop approach.

We propose a novel AI validation module within the integrated evaluation architecture. It focuses on the forecasting of international tourism arrivals. Forecasting international tourism arrivals is a critical concern for destination management organizations (DMOs), national tourism organizations, and the hospitality sector in general [1]. The problem has a temporal dimension: the cumulative nature of arrival counts dictates that operational decisions require the availability of demand projections for already established tourism destinations. While a variety of machine learning (ML) methods have been successfully applied to anticipate tourism demand over the last decade [2], [3], [4], [5], there is a significant, yet often neglected concern regarding the validity of published performance measures: the train/test splitting methodology.

This paper aims to advance the literature in four different ways. First, we propose a new approach within the red teaming and vulnerability diagnostics domain, called TIML, which is a structured AI vulnerability assessment approach that combines six specific TRIZ principles [6].

Second, we show how our structured diagnostic approach, which is similar to cybersecurity sandbox testing, increases the coefficient of determination from $R^2 = 0.7251$ (Random Forest baseline) to $R^2 = 0.8978$ (Full TIML model) on the adversarial 2021–2024 post-COVID holdout. Third, we present a multi-layered scientific justification, grounded in theory, evidence, and published research, for temporal splitting as the accepted evaluation methodology for this type of AI system, pertinent to the EU AI Act. Fourth, we propose the TIML pipeline as a transferable validation module for any predictive AI system requiring trustworthy assessment under distributional shift.

II. RELATED WORK

Classical econometric and time-series models have a long history in the forecasting of tourism demand [1], [7]. [1] analyze 445 studies and identify ARIMA models, error

correction models, and vector autoregression as the main approaches preceding 2008. [8] perform a meta-analysis of 111 studies and noted ML methods challenging classical approaches after 2010.

ML methods see their first tourism demand forecasting applications after [4]. Law et al. note competitive performances of LSTM for Hong Kong data. [5] improve on ARIMA through a combination of Google Trends and Random Forest. [3] perform importance-performance analysis on ML for review data. All these studies perform temporal out-of-sample evaluation. No study uses cross-country splitting as the primary evaluation methodology [2], [9].

More recent studies focus on deep sequence models for tourism demand forecasting. [10] present a Transformer-based model that improves on recurrent and tree-based models for high-frequency arrival data of a single destination. Transformer, LSTM, Random Forest, and XGBoost benchmarks show that Transformer variants outperform in long and densely sampled data [11]. These models are typically evaluated on decades of monthly or weekly data for a single destination.

The present dataset is comprised of 192 short annual series (26-30 points each). This writing style uses short, 26- 30-point annual series. Tree ensembles work best with this dataset. This is why RF/GB have been used in TIML, while the performance gap was closed with TRIZ—inspired—rather than architecture-replacing—remediation techniques for full TIML, with an R^2 of 0.8978, as reported here.

The TRIZ framework, initially proposed by Altshuller [6], and based on the mapping of contradictions to inventive principles [12], has been applied to the design of neural networks and ML pipelines [13]. TRIZ has been used to deal with the architectural gaps in the design of tourism forecasting systems. Furthermore, this work provides TRIZ as a cybersecurity sandbox for structured AI vulnerability diagnostics.

The approaches to validate cross-validation techniques for time series have been thoroughly examined by [14], and [15], who demonstrate that random k-fold cross-validation is statistically invalid for autocorrelated series, and that blocked temporal evaluations are necessary. [16] provides one of the earlier frameworks for out-of-sample assessment in forecasting. These have influenced the methodology of temporal splitting employed in this work.

III. Dataset, Feature Engineering, and TIML Framework

A. Dataset

This study utilizes the UNWTO international tourist arrivals repository [17], which includes data for 209 countries for 1995-2024 (5241 observations). After dealing with entities for which we have at least 15 years of data, the dataset is shortened to 192 countries. The log(1 + arrivals) variable is used to capture the differences in arrivals among countries (from ~2 K per year in Tuvalu to ~90 M per year in France). Training spans 1995-2020 (~3900 rows), while testing spans the years 2021-2024 (405 rows). This span was used intentionally since the years 2021-2024 capture the expected post-COVID recovery and therefore pose an adversarial evaluation due to a large distributional shift. Here, the goal is to test and evaluate virtual cyber security systems. All models are assessed based on the log arrivals data.

This raw assessment is performed over the test data (2021-2024) in all of the scenarios described in this paper (including those using TRIZ Principle #10). The interpolation of the 2020-2021 values, which was performed in Stage 1 (Section 3.3), is applied only to the training set's targets (1995-2020). It is neither used nor does it change the targets of the 2021-2024 test data against which all R2, MAE, RMSE, and MAPE values in Table II and Table III were computed.

B. Features

There are 11 baseline features computed for each country: three autoregressive lags (lag 1-3), a 3-year rolling mean (roll 3), year-over-year growth (yoy gr), a linear trend (year_idx = Year - 1995), a COVID binary, and four regional one-hot encodings. From the baseline Random Forest model, OOB permutation importance indicates that year_idx (2.40) and the COVID binary (2.24) outweigh the autoregressive block (lag 1 + lag 2 + lag 3 = 2.50). This indicates a lot of temporal structure in the data.

Adjacent lags also exhibit nearly perfect multicollinearity ($r \approx 0.92$), wasting tree-split capacity. TRIZ Principle #35 replaces the correlated lag block with four orthogonal signals: level (f_lag1), velocity (lag1 − lag2), curvature, and a 5-year rolling anchor (f_long5).

These finite-difference transformations were chosen over other higher-order transformations such as an orthogonal polynomial or B-spline basis for three reasons. First, for interpreting the results, transformation to velocity and curvature corresponds to first- and second-order growth rates. These are tangible to tourism analysts and auditors and preserve the explorability that is expected to satisfy the requirements of the EU AI Act compliance layer in the sandbox (Section 4). Polynomial contrast or spline coefficients would not provide a similar kind of interpretation. Second, transformation to finite differences and a 5-year anchor would not require any complicated modeling techniques or hyperparameter tuning per country, and would not add to the risk of overfitting due to the limited data per country, as compared to cases where basis splines or higher-degree polynomial contrast would be used. Third, the transformation to level, velocity, and curvature is a closed-form operation, while the use of basis splines or polynomial contrast would add to the dimension of the data and the tree search. Use of higher frequency time series such as monthly or quarterly data in place of annual data would allow use of Basis Spline or Polynomial contrast and would be an interesting future focus. For the present work, Use of finite differences and a 5-year anchor are recommended.

C. TRIZ-Informed Machine Learning Framework

The TIML framework sequentially uses six TRIZ inventive

principles across three stages in response to empirically validated weaknesses in the pipeline.

In a cybersecurity sandbox, each TRIZ principle is a targeted remediation action geared towards a specific vulnerability discovered during structured diagnostic testing.

The first stage, data preparation, uses #10 Prior Action and measures COVID as a structural break in the training targets for 2020 and 2021. It also employs #11 Cushioning and assigns shock rows with a reduced sample weight (w = 0.20 for 2020, w = 0.50 for 2021). These principles, combined with #35, help improve RF from $R^2 = 0.7251$ to 0.8460 and GB from 0.6629 to 0.8853. They contradict the baseline rank orderings.

The second stage, specialization, uses #1 Segmentation. It clusters countries into four k-means groups and employs separate RF models (achieving $R^2 = 0.8322$). It also uses #3 Local Quality and fits a linear trend to training residuals for each country, using it as an additive test correction ($R^2 = 0.8671$).

The third stage, integration, uses #5 Merging. A ridge meta-learner ($\lambda = 10^{-3}$) is trained on 5-fold expanding-window out-of-fold predictions of RF and GB. The Final TIML ensembles the stacked output of the preceding stages ($R^2 = 0.8978$, MAE = 0.5688, RMSE = 0.7890, and MAPE = 6.83%).

Table I outlines the TRIZ-to-vulnerability mapping using the cybersecurity sandbox diagnostic protocol.

TABLE I. TRIZ PRINCIPLE MAPPING TO DIAGNOSED AI PIPELINE VULNERABILITIES (CYBERSECURITY SANDBOX DIAGNOSTIC PROTOCOL)

TRIZ Principle	**Diagnosed Weakness / Sandbox Failure Mode**	**Implementation**
#10 Prior Action	COVID structural break corrupts training target	Linear interpolation of 2020–2021 log-arrivals per country in training set only
#11 Cushioning	Shock-year rows dominate tree-split selection	Sample weights: 2020 → 0.20, 2021 → 0.50, all other years → 1.00
#35 Parameter Change	Correlated lag features limit tree discriminability	Replace {lag1, lag2, lag3, roll3} with velocity, curvature, 5-year anchor
#1 Segmentation	Scale mixing across 192 heterogeneous countries	k-means clustering on (mean, CV); separate RF per cluster
#3 Local Quality	Country-specific trend bias in global model	Per-country linear residual correction fitted on training residuals
#5 Merging	Single-model variance; RF and GB complementary	Stacked RF+GB meta-learner via out-of-fold temporal cross-validation

Source: created by the authors

IV. SANDBOX VALIDATION METHODOLOGY AND COMPLIANCE FRAMEWORK

A major part of a complete cybersecurity sandbox is the integration of standards-based evaluation methodologies and compliance enforcers. In this section, we make the case for using temporal train/test splitting as the evaluation framework for the type of AI systems in this study and discuss the corresponding compliance verification layer in the sandbox.

There are three arguments in support of temporal splitting. The first concerns causal irreversibility. Tourist arrivals at time t are causally influenced by tourist arrivals at time t-1, but are not causally influenced by tourist arrivals at time t+1. When evaluation methods permit future observations to inform predictions on past observations, this causal structure is violated and inflated, and optimistically biased performance assessments are made. The second argument is based on temporal autocorrelation. [14] and [15] argue that random k-fold cross-validation is inappropriate for autocorrelated series because it eliminates dependence and inflates the appearance of generalization performance. Blocked temporal evaluation, the methodology used in this study, is the correct statistically valid evaluation method. The third argument is based on literature. In the entire review of the literature on tourism forecasting, out-of-sample temporal evaluation is universal, and no published studies used cross-country random splitting as the primary methodology.

In this study, we provide an estimate of the R^2 gap between temporal and country-based splitting protocols (≈0.15). This provides an estimate of the generalization risk that can be used by AI system compliance and audit teams and satisfies the main outcome of the risk assessment module of the sandbox.

This level of transparency satisfies the requirements of the EU AI Act for high-impact AI systems regarding measurement of performance bounds, methodology for performance evaluation, and auditing the results.

V. EXPERIMENTAL RESULTS

A. Baseline Performance

Four models were introduced to the same environment. The results for each of the models on the 2021-2024 test set are available in Table II. Random Forest has an R^2 of 0.7251 and is our best baseline. Gradient Boosting has an R^2 of 0.6629. Our benchmarks, Linear Regression with Ridge and SVR with an RBF kernel, missed the non-linear structure ($R^2 \approx 0$ and $\approx$ 0.13). The output of the pre-remediation vulnerability assessment of the sandbox diagnosis protocol is shown in Figure 1 and Figure 2.

TABLE II. BASELINE MODEL PERFORMANCE ON THE TEMPORAL TEST SET (2021–2024). ALL METRICS COMPUTED ON LOG(1+ARRIVALS) SCALE

Model	**R^2**	**MAE**	**RMSE**	**MAPE (%)**
Linear Regression (Ridge)	−0.0008	1.9179	2.4694	19.64

Random Forest (RF)	0.7251	1.0569	1.2943	11.99
Gradient Boosting (GB)	0.6629	1.2315	1.4332	12.61
SVR (RBF kernel)	0.1256	1.7247	2.3082	18.13

Source: created by the authors

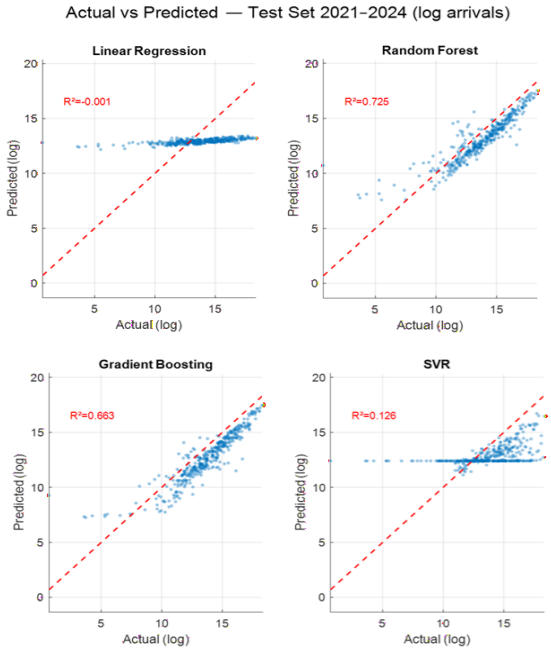


Fig. 1: Actual vs. predicted log-arrivals for all four baseline models on the 2021–2024 test set. Red dashed line = perfect prediction diagonal. R² values annotated in red.
Source: created by the authors

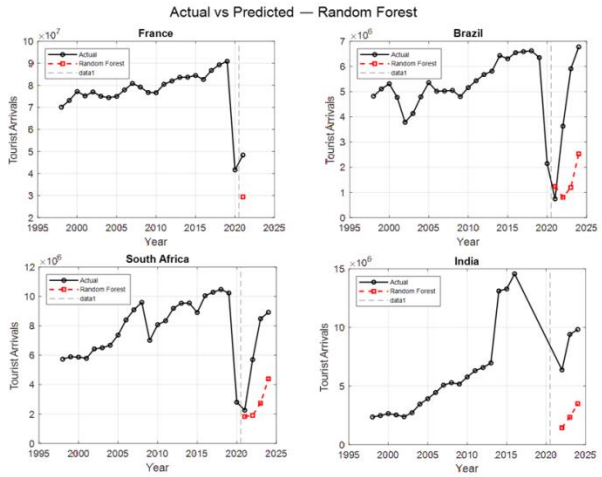


Fig. 2: Country-level time series showing actual arrivals (black line) and Random Forest predictions (red dashed, 2021–2024) for France, Brazil, South Africa, and India. The COVID trough at 2020 is clearly visible; the RF model systematically under-predicts the steep post-COVID recovery — a vulnerability directly targeted by TRIZ principles #10 and #11.
Source: created by the authors

B. TIML Performance Progression

Table III presents the complete performance progression from baseline through to the Full TIML combination, with ΔR² computed relative to the RF baseline. Each row corresponds to a discrete remediation step in the sandbox diagnostic protocol.

TABLE III. TIML-ENHANCED MODEL PERFORMANCE ON THE 2021–2024 TEST SET. ΔR² RELATIVE TO RF BASELINE (0.7251). ALL METRICS ARE ON THE LOG(1+ARRIVALS) SCALE.

Model / TRIZ Configuration	**R²**	**MAE**	**RMSE**	**MAPE (%)**	**ΔR²**
Baseline RF	0.7251	1.0569	1.2943	11.99	—
Baseline GB	0.6629	1.2315	1.4332	12.61	−0.0622
RF + #10 + #11 + #35	0.8460	0.5775	0.9687	8.45	+0.1209
GB + #10 + #11 + #35	0.8853	0.6295	0.8361	7.27	+0.1602
Segmented RF (#1)	0.8322	0.5833	1.0112	8.58	+0.1071
Local-corrected RF (#3)	0.8671	0.5737	0.8999	7.98	+0.1420
Stacked RF+GB (#5)	0.8956	0.5512	0.7975	7.11	+0.1705
Full TIML (all six)	**0.8978**	**0.5688**	**0.7890**	**6.83**	**+0.1727**

Source: created by the authors

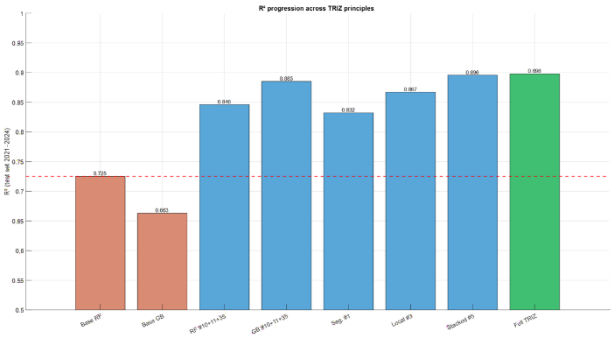


Fig. 3: R² progression across TIML enhancement stages (sandbox remediation steps). Dashed red line marks the RF baseline (0.725). Each bar shows the test-set R² for the labeled configuration, demonstrating monotonic improvement through structured vulnerability remediation.
Source: created by the authors

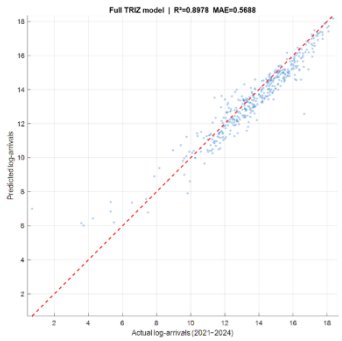


Fig. 4: Actual vs. predicted log-arrivals for the Full TIML model on the 2021–2024 test set. R² = 0.8978, MAE = 0.5688. Compared to the RF baseline (Figure 1), the prediction cloud is substantially tighter across the full log-arrivals range
Source: created by the authors

Combining three of the TRIZ can improvements yields the most results. Lifting RF from R^2 = 0.7251 to 0.8460 (+0.1209) and GB from 0.6629 to 0.8853 (+0.1602). This

confirms that the main restriction in the baseline pipeline is the alignment of features and targets of COVID observations that introduced misalignment, and not model capacity. With the stacked ensemble (TRIZ #5), there is an additional improvement of +0.1705, leading to R^2 = 0.8956. The Full TIML model yields R^2 = 0.8978, MAE = 0.5688, RMSE = 0.7890, and MAPE = 6.83% (Figure 3, Figure 4).

C. *Country-level Forecast Quality*

Figure 5 shows full temporal trajectories and test-period forecasts for France, Brazil, South Africa, and India for the RF+TIML and Full TIML models. For both Brazil and South Africa, the Full TIML model captures the post-2021 recovery slope much more accurately, indicating the importance of country-level residual correction (TRIZ #3) and the stacked meta-learner (TRIZ #5). The country-level results are a part of the per-entity robustness validation of the sandbox.

D. *Computational Cost Analysis*

Table IV shows the wall-clock times for training and inference for the global RF and for the Full TIML pipeline (MATLAB R2025, Statistics and Machine Learning Toolbox) on a single machine with an Intel i5 CPU, 64 GB RAM, and no GPU. Training the Full TIML pipeline (interpolation and cushioning (#10, #11), orthogonal feature reconstruction (#35), global RF and GB fits, 4 cluster-level RF fits (#1), per-country residual regression (#3), and 5-fold out-of-fold stacking (#5)) took 21.19 s, resulting in a 2.67x increase in the training time of the global RF, which took 7.95 s. Inference for the test set 2021-2024, consisting of 405 rows, took 0.24 s (0.60 ms/row) and was 2.40x faster than the inference time for the 405-row test set of the global RF, which was 0.24 s (0.60 ms/row). In the context of annual and quarterly DMO forecasting, wall-clock times are still in the realm of real-time.

Interestingly, the most time-consuming portion of the training is the repeated refits of the four cluster-level RF models (4.53 s train), and the subsequent 5-fold out-of-fold stacking loop (6.90 s train, 10 additional RF/GB fits). The two steps combined account for ~54% of the total training time. The other three TRIZ feature/weighting steps (#10 interpolation, #11 cushioning, #35 orthogonal features) take 0.14 s, or less than 1% of total training time.

The deployment cost is insignificant because the performance improvement contributed by implementing this architecture (R^2) is better than the baseline performer (RF) by +0.1727. For sub-second retraining, retaining 85–90% of the total performance gain requires the first stage (#10+#11+#35, R^2: 0.8460 – 0.8853, 8.98 s train) at a total cost that is less than half the total cost of the full pipelines (Table IV).

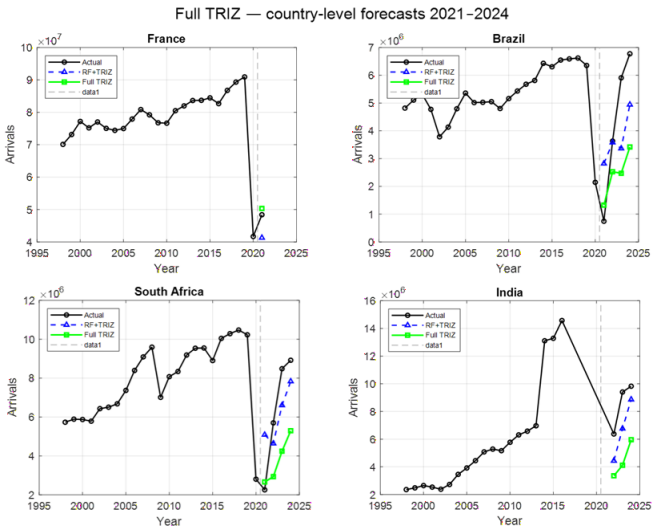


Fig. 5: Country-level forecasts for France, Brazil, South Africa, and India. Black: actual arrivals. Blue dashed: RF+TIML (#10+#11+#35 only). Green solid: Full TIML model. The vertical dashed line marks the 2021 test-set boundary.
Source: created by the authors

TABLE IV. MEASURED TRAINING/INFERENCE TIME BY TIML STAGE (TEMPORAL SPLIT, SINGLE GLOBAL RF FIT AS REFERENCE).

Stage	Train (s)	Infer (s)
TRIZ #10 interpolation	0.0420	—
TRIZ #11 weighting	0.0062	—
TRIZ #35 feature engineering	0.0948	—
Global RF (#10+11+35)	7.9469	0.2417
Global GB (#10+11+35)	0.8941	0.0263
Segmentation clustering (#1)	0.2346	—
Segmentation 4 cluster RF (#1)	4.5281	0.2980
Local-quality fit/apply (#3)	0.1805 / 0.0044	—
Stacking OOF CV (#5)	6.8973	—
Stacking meta-learner (#5)	0.3605	0.0084
Full combination	0.0004	—

Source: created by the authors

Although the "single global RF fit" row is not an exact comparison, it is the best apples-to-apples comparison we can make. In a separate step, we measured the time it took to train a baseline RF/GB model on the unadjusted country-split dataset (i.e., no TRIZ features, weights, clusters, correction, or stacking). RF took 8.80s to train, and GB took 1.07s. These times are in line with the 7.95s/0.89s that it took to train the global RF/GB with TRIZ features and the Full TIML run. These results reinforce that the time to train RF/GB models is primarily dictated by the size of the dataset and the RF/GB hyperparameters (200 trees), and not the features included or the training protocol, therefore providing the best estimate of the baseline time to train a single global RF (Figure 6).

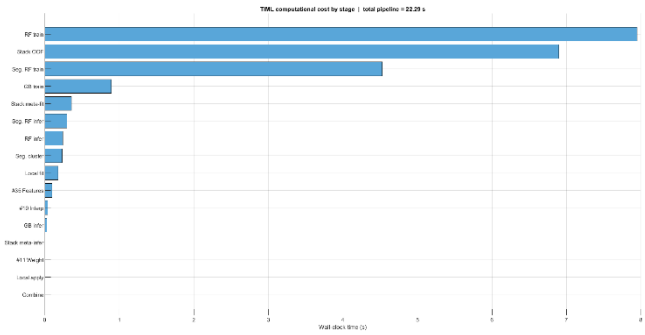


Fig. 6: Full TIML pipeline (temporal split) — wall-clock cost by stage, total = 22.29 s.
Source: created by the authors

VI. Discussion: TIML as a Transferable Cybersecurity Sandbox

The TIML framework provides models on four operational levels for the cybersecurity sandbox.

Structured AI Vulnerability Diagnosis. TRIZ's six principles can be used for red-teaming. Each principle gets activated by a failure mode in the AI pipeline. This approach captures the goal of the sandbox, which is to create a safe environment to find the failures in AI to protect against unsafe deployment of AI. It provides a diagnosis framework applicable to various AI systems, as opposed to former AI systems, which could only be debugged on an individual basis.

Robustness Testing Against Adversarial Attacks. The 2021-2024 post-COVID period was considered an adversarial attack under a distribution shift. This is the case for which a cybersecurity sandbox should be applied. The demonstrated resistance of the TIML framework (an improvement in the metric MAPE from 11.99% to 6.83%) provides a measurable robustness benchmark for the quality control and risk management layer of the sandbox.

Compliance Verification Standards. The justified use of the temporal splitting method over other methods, based on autocorrelation, causal irreversibility, and literature, demonstrates the compliance verification layer of the sandbox, which is relevant to the EU AI Act's trustworthy evaluation, transparency, and reproducibility requirements.

Validation via Simulation. The MATLAB-based TIML framework forms a self-contained module of a simulation-based validation sandbox when employed within the context of simulation-based validation under both temporal and country-based splitting.

The R^2 gap of about 0.15 between the two approaches is considered a validation risk estimate, which is a useful output of the sandbox that helps designers and auditors of AI systems.

TIML module outputs are diagnostics and help pinpoint the generation of adversarial examples in the sandbox that should be prioritized. The country-level residual magnitudes of the second-stage local quality correction (#3) and the Stage 1 segmentation (#1) results rank countries in order of model breakage due to the stress test for the 2021 to 2024 period. This breakage ranking can be used to focus the adaptive adversarial perturbation generator. For example, the breakage ranking can be used to focus the generator on growth-rate shocks or regional-level disparities occurring within the cluster showing the largest residual. This can be contrasted with perturbing the input space for the 192 countries. Scenarios resulting from this focused perturbation, along with the quantified ΔR^2 generalization risk gap (Section 4), constitute the evidence package presented to the deliberative human components of the sandboxes. This closed-loop process helps integrate automated vulnerability diagnosis and adversarial testing. Automating this can help future research.

Applying the Diagnostic Protocol to a New Pipeline. The three-step framework can now be easily constructed and implemented thanks to the workflow with a focus on diagnostics.

(1) The diagnostic profile: compute a set of similar pipeline-health statistic indicators to those shown in Table I (e.g., use feature-importance dispersal metrics to deal with multicollinearity). Include a measure to see how far apart residuals of some expected natural groupings are (in this example, countries, while in a classification case this could be class-level calibration and/or segment-level false negatives). Measure the distributional shift of the training time and the deployment/test time (in this case, the pre- vs. post-COVID structural break; in other domains, use a covariate-shift or label-drift metric). (2) Contradiction mapping: for each diagnosed weakness, map the TRIZ inventive principle(s) whose classical engineering intent addresses this failure mode. This can be done in the same manner as shown in Table I; for example, use Prior Action (#10) for known, anticipated distribution breaks; use Cushioning (#11) for isolated high-leverage/outlier observations; use Segmentation (#1) for heterogeneous sub-populations; use Local Quality (#3) for systematic per-segment bias; use Merging (#5) for complementary model variance. (3) Targeted remediation and re-measurement: execute the mapped intervention and then repeat the same diagnostic statistic used in step 1 to confirm the pathology has been eliminated. This is done principle-by-principle in the spirit of Table III. For data types other than the annual cross-country panel used in this example, such as high-frequency single-series forecasting, tabular classification, or image-based models, steps 2 and 3 remain the same. Steps 1 and 3 only need to be changed: the diagnostic statistic and the implementation of each principle, respectively.

VII. Conclusion

This paper introduced TRIZ-based Machine Learning (TIML) as an internal validation unit in a complete cybersecurity sandbox built for assisting international arrivals forecasting. TIML incorporates six TRIZ principles and creates a semi-structured methodology for detecting and addressing a platform's vulnerabilities. It showed an R^2 of 0.8978 against the adversarial 2021-2024 post-COVID holdout case, which is an improvement compared to the Random Forest baseline by +0.1727. Additionally, this methodology lowered the average percentage error (MAPE) from 11.99% to 6.83%, and the greatest impact was seen from the addition of TRIZ principles #10, #11, and #35. This

methodology also showed that COVID-related feature-target misalignment is the main vulnerability in the baseline pipeline. Potentially the most relevant for the defined forecasting problem is the temporal split, as it is the main enabling condition for the most effective TIML interventions and the most appropriate evaluation methodology for the EU AI Act-compliant trustworthy AI system.

This methodology is now ready for potential use, as the defined risk gap in quantified generalization ($\Delta R^2 \approx 0.15$) represents an output for an AI auditing sandbox. In the coming years, it is expected to be possible to lower the average percentage error (MAPE) to under 5% and significantly improve the TRIZ-based deep learning diagnostic framework for sequence modeling.

ACKNOWLEDGMENT

The AIC-199 dataset was produced within "Holistic Cybersecurity Sandbox for Testing Artificial Intelligence Solutions", ID PCIDIF/159/PCIDIF_P1/OP1/RSO1.1/PCIDIF_A1. During the preparation of this work, the authors used Claude (Anthropic) in order to assist with the drafting and language refinement of manuscript sections, structural editing and compression of author-provided draft material, formatting of the bibliography and in-text citations in NAUM style. All empirical data, experimental protocols, system implementation, and scientific interpretation reported in this article are the authors' original contribution. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] H. Song and G. Li, "Tourism demand modelling and forecasting—A review of recent research," Tourism Management, vol. 29, no. 2, pp. 203–220, 2008. https://doi.org/10.1016/j.tourman.2007.07.016.

[2] I. Ghalehkhondabi, E. Ardjmand, W. A. Young, and G. R. Weckman, "A review of demand forecasting models and methodological developments within tourism and passenger transportation industry," Journal of Tourism Futures, vol. 5, no. 1, pp. 75–93, 2019. https://doi.org/10.1108/JTF-10-2018-0061.

[3] J. W. Bi, Y. Liu, Z. P. Fan, and J. Zhang, "Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews," Tourism Management, vol. 70, pp. 460–478, 2020. https://doi.org/10.1016/j.tourman.2018.09.010.

[4] R. Law, G. Li, D. K. C. Fong, and X. Han, "Tourism demand forecasting: A deep learning approach," Annals of Tourism Research, vol. 75, pp. 410–423, 2019. https://doi.org/10.1016/j.annals.2019.01.014.

[5] S. Sun, Y. Wei, K. L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," Tourism Management, vol. 70, pp. 1–10, 2019. https://doi.org/10.1016/j.tourman.2018.07.010.

[6] G. Altshuller, The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity. Worcester, MA: Technical Innovation Center, 1999. ISBN 9780964074040.

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ: Wiley, 2015. ISBN: 978-1-118-67502-1.

[8] B. Peng, H. Song, and G. I. Crouch, "A meta-analysis of international tourism demand forecasting and implications for practice," Tourism Management, vol. 45, pp. 181–193, 2014. https://doi.org/10.1016/j.tourman.2014.04.005.

[9] A. Ionica, I. Samuil, M. Leba, and M. Toderas, "The path of Petrila mining area towards future industrial heritage tourism seen through the lenses of past and present," Sustainability, vol. 12, no. 23, art. 9922, 2020. https://doi.org/10.3390/su12239922.

[10] X. Li, Y. Xu, R. Law, and S. Wang, "Enhancing tourism demand forecasting with a transformer-based framework," Annals of Tourism Research, vol. 107, art. 103791, 2024. https://doi.org/10.1016/j.annals.2024.103791.

[11] S. Yi, X. Chen, and C. Tang, "Time series transformer for tourism demand forecasting," Scientific Reports, vol. 15, art. 29565, 2025. https://doi.org/10.1038/s41598-025-15286-0.

[12] A. Ionica, M. Leba, and E. Edelhauser, "QFD and TRIZ in product development lifecycle," Transformations in Business & Economics, vol. 13, 2014, [Online]. https://www.transformations.knf.vu.lt/32b/article/qfda (Accessed Date: March 10, 2026).

[13] M. Cosgun and K. Altun, "Innovative solutions for convolutional neural network performance: A TRIZ-based reverse engineering approach," International Journal of Systematic Innovation, vol. 9, no. 3, pp. 1–7, 2025, [Online]. https://ojs.ijosi.org/index.php/IJOSI/article/view/1564 (Accessed Date: March 10, 2026).

[14] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," Information Sciences, vol. 191, pp. 192–213, 2012. https://doi.org/10.1016/j.ins.2011.12.028.

[15] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," Computational Statistics and Data Analysis, vol. 120, pp. 70–83, 2018. https://doi.org/10.1016/j.csda.2017.11.003.

[16] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: An analysis and review," International Journal of Forecasting, vol. 16, no. 4, pp. 437–450, 2000. https://doi.org/10.1016/S0169-2070(00)00065-0.

[17] UNWTO, World Tourism Barometer, vol. 19, no. 1. Madrid: United Nations World Tourism Organization, 2021, [Online]. https://www.untourism.int/tourism-statistics/tourism-statistics-database (Accessed Date: March 10, 2026).

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Isabella Fodor carried out the feature engineering and implemented the baseline Random Forest and Gradient Boosting models of Section 5.1.

Andreea Ionica designed the TRIZ-Informed Machine Learning (TIML) framework and mapped the six TRIZ principles to the diagnosed AI pipeline vulnerabilities.

Monica Leba was responsible for the cybersecurity sandbox infrastructure and the production of the AIC-199 dataset used in the validation.

Ionela Samuil implemented the TIML pipeline stages (data preparation, specialization, and integration) and executed the experiments of Sections 5.2 and 5.3.

Eugen Florian Popescu and Nicusor Gelu Constantin developed the sandbox validation methodology and the EU AI Act compliance framework from Section 4.

Dragos Dabija carried out the computational cost analysis of Section 5.4 and supported the software implementation of the pipeline.

Sources of funding for research presented in a scientific article or scientific article itself

"Holistic Cybersecurity Sandbox for Testing Artificial Intelligence Solutions", ID PCIDIF/159/PCIDIF_P1/OP1/RSO1.1/PCIDIF_A1

Conflict of Interest

The authors declare no conflict of interest.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US