

Pakistan Journal of Criminal Justice
ISSN: 2958-9363 ISSN (E): 2958-9371
Volume 6, No.8, 2026, 193-218
DOI: https://doi.org/10.59075/pjcj.v6i8.329
Journal Homepage: https://journals.centeriir.org/index.php/pjcl

From Algorithmic Anticipation to Normative Accountability: A Multimodal Deep Learning and Big Data Analytics Framework for Early Warning and Pre-emptive Detection of Violent Extremism, with Explainable AI, Graph-Based Radicalization Trajectory Modelling and an Ethical–Legal Governance Architecture for Rights-Compliant Counter-Extremism

Hamza Sajjad[1*], Dr. Muhammad Usman Malik[2], Dr. Zulfqar Khan[3]

[1]National Institute of Intelligence and Security Studies, Quaid-i-Azam University, Islamabad 45320, Pakistan
Email**:** hamzasajjad0333@gmail.com
[2]National Institute of Intelligence and Security Studies, Quaid-i-Azam University, Islamabad 45320, Pakistan
Email: usmanmalik836@gmail.com
[3]National Institute of Intelligence andSecurity Studies, Quaid-i-Azam University, Islamabad 45320, Pakistan,
Email: dirniiss@qau.edu.pk

Abstract

Purpose: Predictive analytics are increasingly proposed for counter-extremism, yet most frameworks optimize anticipation while treating accountability as an afterthought. This study designs and empirically tests an integrated framework that couples algorithmic early warning with an ethical–legal governance architecture, using Pakistan ranked first in the Global Terrorism Index 2026 as the test case.

Design/methodology/approach: A real-data workbook (PIPS 2006–2025, PICSS, CRSS, SATP, IEP-GTI, DataReportal, and 24 coded legal instruments) feeds four modules: (i) an econometric core (bootstrap ADF, KPSS, Zivot–Andrews, Bai–Perron multiple breaks, Hamilton Markov-switching, negative-binomial log-linear INGARCH, Johansen/Engle–Granger cointegration, VAR with bootstrapped generalised impulse responses, unobserved-components Kalman smoothing and critical-slowing-down indicators); (ii) deep learning (from-scratch LSTM and deep multilayer perceptron ensembles, in level and residual-learning form) evaluated by rolling-origin forecasting and Diebold–Mariano tests; (iii) explainable AI (exact Shapley values, integrated gradients, permutation importance, counterfactuals); and (iv) a graph-based radicalization-trajectory model (maximum-entropy actor province edges, simplified graph convolution, empirical-Bayes actor growth).

Findings: Violence follows four regimes (breaks in 2010, 2015 and 2020). A Markov-switching filter assigned a 0.94 real-time probability to an escalation regime in 2021, when attacks were still only 207. The NAP/Zarb-e-Azb regime is associated with a 62.5% lower expected attack intensity and the post-Kabul regime with a 57.2% higher intensity. Fatalities scale almost proportionally with attacks (elasticity 0.95). Deep learning ensembles did not outperform parsimonious baselines out of sample (all Diebold–Mariano $p > 0.10$); damped trend smoothing achieved the lowest error. Graph modelling shows nationalist insurgency expanding fastest (Baloch groups: log-growth 1.47 per year).

Originality/value: The paper converts these results into a rights-compliant governance architecture aggregate-only outputs, mandatory benchmark disclosure, human-in-the-loop thresholds and audit trails and shows that accountability must include *epistemic* accountability for model performance.

Keywords: violent extremism; early warning; deep learning; explainable AI; graph neural networks; Markov switching; INGARCH; counter-terrorism governance; human rights; Pakistan

OPEN ACCESS
© 2026 The Authors, Published by Center of Innovation in Interdisciplinary Research (CIIR).
This is an Open Access Article under the Creative Common Attribution Non-Commercial 4.0

Article History: Received: 12-08-2026 Accepted: 02-09-2026 Published: 16-10-2026

Corresponding Author's Email: hamzasajjad0333@gmail.com

JEL Classification: C22, C25, C45, C53, D74, H56, K42

Introduction

Violent extremism places states in a tension. On one side is the strategic demand to anticipate attacks before they occur; on the other is the constitutional and human-rights obligation not to punish, surveil or stigmatize people for what they might do. Advances in deep learning, big data analytics and graph learning have sharpened the first demand by promising earlier, finer-grained and more automated warning (Hegre et al., 2019; Mueller & Rauh, 2018). They have not, by themselves, resolved the second. Predictive policing research shows that algorithmic forecasts trained on enforcement data can reproduce and amplify existing bias through feedback loops (Ensign et al., 2018; Lum & Isaac, 2016; Richardson et al., 2019), and the most recent European legislation now prohibits crime-risk prediction based solely on profiling (Regulation (EU) 2024/1689, Art. 5). An early-warning system for violent extremism that is accurate but unaccountable is therefore neither lawful nor, in the long run, effective.

Pakistan is an important test case. After a decade-long decline from the 2009 peak of 2,586 terrorist attacks recorded by the Pak Institute for Peace Studies (PIPS), violence reached a trough of 146 attacks in 2020 and then rose for five consecutive years to 699 attacks and 1,034 deaths in 2025 a 379% rise in attacks in five years. In the Global Terrorism Index 2026, Pakistan was ranked the most terrorism-affected country in the world for the first time, with a score of 8.574 (Institute for Economics & Peace [IEP], 2026). The resurgence is multi-actor: Tehreek-e-Taliban Pakistan (TTP) accounted for the majority of terrorism deaths, while Baloch nationalist groups, including the Balochistan Liberation Army (BLA), expanded sharply. The country also has a dense counter-terrorism legal architecture, from the Anti-Terrorism Act 1997 to the Prevention of Electronic Crimes Act 2016 and its 2025 amendment, which makes the governance question concrete rather than abstract.

This paper asks three questions. **RQ1:** Can a multimodal analytics framework that combines advanced econometrics, deep learning and graph learning detect regime shifts in extremist violence early enough to be useful? **RQ2:** Which signals drive those warnings, and can they be explained in a form that decision-makers and oversight bodies can audit? **RQ3:** What ethical–legal architecture is needed so that such a system remains rights-compliant, and how should its design respond to the empirical limits of prediction?

The paper contributes in four ways. First, it builds and tests a four-layer framework data ingestion, an anticipation engine, explainability and evaluation, and governance on a fully sourced real-data workbook rather than simulated data. Second, it applies a battery of demanding econometric techniques rarely combined in terrorism studies: sieve-bootstrap unit-root inference, Bai–Perron multiple structural breaks, a Hamilton Markov-switching model with Kim smoothing, a negative-binomial log-linear INGARCH count model with regime covariates, Johansen and Engle–Granger cointegration, a VAR with bootstrapped generalised impulse responses, an unobserved-components model estimated by Kalman filtering, and critical-slowing-down indicators tested against surrogate data. Third, it implements deep learning models from first principles, explains them with exact Shapley values and integrated gradients, and benchmarks them honestly against parsimonious baselines. Fourth, it translates the results into a governance architecture anchored in Pakistani law and international instruments, including the ICCPR, UN Security Council Resolution 2396, the UN Plan of Action to Prevent Violent Extremism, the EU AI Act and the Council of Europe Framework Convention on Artificial Intelligence.

The central empirical message is two-sided. Regime-aware econometric filters provided timely and interpretable warning of Pakistan's 2021 turning point. Deep learning ensembles, by contrast, did not outperform simple baselines on twenty annual observations. We argue that this is not a failure of the framework but its most important governance output: *normative accountability* must include *epistemic*

accountability the obligation to disclose when complex models add no predictive value over transparent ones. The remainder of the paper is organised as follows. Section 2 reviews the literature; Section 3 sets out the theoretical framework and hypotheses; Section 4 describes data and methods; Section 5 reports results; Section 6 presents the governance architecture; Section 7 discusses implications; and Section 8 concludes.

Literature Review

Forecasting Political Violence and Terrorism

Quantitative forecasting of political violence has moved from explanatory regression toward out-of-sample prediction. Ward et al. (2010) showed that variables with significant coefficients often add little predictive power, and Schrodt (2014) warned against overfitting and the neglect of out-of-sample validation. Large-scale systems such as ViEWS produce monthly conflict forecasts and evaluate them with proper scoring rules (Hegre et al., 2019, 2021), while Mueller and Rauh (2018) showed that topic models of newspaper text improve conflict prediction. Cederman and Weidmann (2017) cautioned that forecasting armed conflict is limited by the rarity and contingency of onset events. In terrorism research, time-series studies have documented persistence, cycles and structural change in attack series (Enders & Sandler, 2005; Enders et al., 2011), and the heavy-tailed distribution of attack severity (Clauset et al., 2007). Self-exciting point-process models, borrowed from seismology, capture contagion between events (Mohler et al., 2011). These studies show that terrorism series are persistent, regime-dependent and overdispersed properties that any early-warning model must respect.

Deep Learning, Big Data and Online Radicalization

Deep recurrent architectures, especially the long short-term memory network (Hochreiter & Schmidhuber, 1997), dominate sequential forecasting in data-rich settings. Graph neural networks extend deep learning to relational data (Kipf & Welling, 2017), and simplified graph convolution shows that much of their power comes from feature propagation over a normalized adjacency matrix (Wu et al., 2019). In extremism research, machine learning has been used to detect extremist content and predict adoption of extremist messages online (Ferrara et al., 2016), to map the online ecology of pro-ISIS aggregates (Johnson et al., 2016) and to trace radical posting behaviour over time (Scrivens et al., 2020). Yet the empirical evidence on the internet's causal role in radicalization remains mixed (Conway, 2017; Gill et al., 2017). The gap is between models that classify content and models that anticipate violence at the level where policy is made.

Radicalization as a Trajectory and a Network

Theories of radicalization emphasize pathways rather than profiles. Moghaddam's (2005) staircase model and McCauley and Moskalenko's (2008) mechanisms of political radicalization describe progressive movement through stages, while Borum (2011) stresses that there is no single terrorist profile. Sageman (2004) showed that militant networks grow through social ties. These insights support modelling radicalization at the level of groups and their spatial reach who operates where, and how fast—rather than at the level of individual risk scores. In Pakistan, the literature documents the TTP's reconstitution after the August 2021 Taliban takeover in Kabul, the rise of Baloch insurgency and the shifting role of Islamic State–Khorasan (Fair, 2014; Jadoon, 2021).

Explainability and Accountability

Explainable AI provides tools to attribute predictions to inputs: Shapley-value methods (Lundberg & Lee, 2017; Shapley, 1953), local surrogates (Ribeiro et al., 2016), integrated gradients (Sundararajan et al., 2017), graph explainers (Ying et al., 2019) and counterfactual explanations (Wachter et al., 2018). Rudin (2019) argues that for high-stakes decisions, inherently interpretable models should be preferred over post-hoc explanations of black boxes. Doshi-Velez and Kim (2017) call for rigorous evaluation of interpretability. In policing, scholars have documented how big-data tools reshape surveillance (Brayne, 2017; Ferguson, 2017) and how randomized trials are needed to establish their effects (Mohler et al., 2015). Critical scholarship on countering violent extremism warns of securitization of communities (Kundnani & Hayes, 2018).

Research Gap

Three gaps follow. First, few studies combine regime-switching econometrics, count time-series models and deep learning within one early-warning design, and fewer benchmark deep models against parsimonious baselines with formal tests. Second, graph approaches to radicalization rarely address the problem that actor-to-location ties are partially observed. Third, governance is usually discussed separately from modelling, so that the legal analysis cannot respond to what the models can and cannot do. This paper addresses all three by making the governance layer a function of the empirical evaluation.

Theoretical Framework and Hypotheses

We conceptualize early warning as a two-stage problem: *algorithmic anticipation*, the statistical detection of changes in the latent intensity of violence, and *normative accountability*, the legal and ethical constraints that govern how such detection may be used. Drawing on critical-transition theory (Scheffer et al., 2009), we model extremist violence as a system that can move between persistent regimes escalation and de-escalation driven by actor capability, state response and external shocks. Regime theory implies that warning signals should appear as changes in the parameters of the data-generating process (breaks, regime probabilities, latent slope), not only as changes in the level of violence.

From the radicalization-pathway literature (McCauley & Moskalenko, 2008; Moghaddam, 2005) we derive a relational layer: groups (nodes) project violence into territories (edges), and expansion appears as growing edge weights and new edges. From the accountability literature (Rudin, 2019; Wachter et al., 2018) we derive a governance principle: the legitimacy of a model depends on its demonstrated value over simpler, transparent alternatives and on the contestability of its outputs. The framework (Figure 1) operationalizes these ideas. We test five hypotheses:

- **H1 (Regimes):** Pakistan's terrorist-attack series is characterized by multiple structural breaks and persistent escalation/de-escalation regimes.
- **H2 (Policy and shocks):** The post-2014 National Action Plan (NAP) and Zarb-e-Azb regime is associated with lower attack intensity, and the post-August-2021 regime with higher intensity, after controlling for count persistence.
- **H3 (Lethality):** Fatalities scale proportionally with attacks in the long run (unit elasticity, cointegration).
- **H4 (Deep learning):** Deep learning ensembles provide more accurate one-step-ahead forecasts than parsimonious statistical baselines.

- **H5 (Graph trajectories):** Actor growth trajectories are heterogeneous, with nationalist insurgency expanding faster than religiously inspired militancy in the post-2020 period.

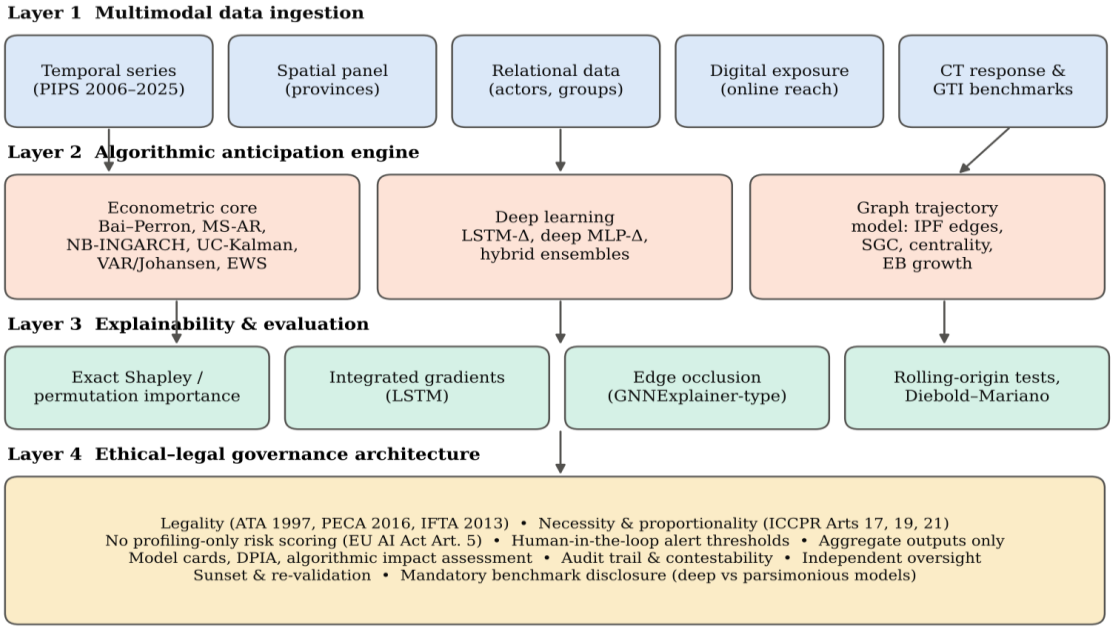


Figure 1. The proposed four-layer framework. Layer 2 combines an econometric core, deep learning and graph learning; Layer 3 explains and evaluates; Layer 4 constrains use through legal and ethical requirements. Source: Authors.

Data and Methodology

Data

The analysis uses a real-data workbook compiled in November 2026 from published sources; it contains no simulated values. Table 1 summarizes the components. The core series is the PIPS annual record of terrorist attacks, persons killed and injured in Pakistan for 2006–2025 ($T = 20$). Twelve years are verified against PIPS releases and press coverage; eight years (2008 and 2010–2016) are flagged in the workbook as recalled from earlier PIPS reports and not re-verified. We therefore report a dedicated robustness analysis for these years (Section 5.10). Because PIPS, PICSS, CRSS, SATP and IEP use different definitions—for example, CRSS and PICSS fatalities include militants killed in counter-terrorism operations—we never mix sources within a series.

Table 1. Data Components, Sources and Framework Roles

Component	Source	Coverage	Framework role
Annual attacks, killed, injured	PIPS Pakistan Security Reports	2006–2025 (T = 20)	Temporal forecasting; econometric core
Provincial attacks and deaths	PIPS PSR 2024, 2025; CRSS 2025	5–6 regions × 2 years	Graph nodes; spatial risk

Component	Source	Coverage	Framework role
Actor typology	PIPS; IEP-GTI 2024–2026	3 classes; TTP, BLA	Graph nodes/edges; trajectories
Combat-death composition	PICSS, CRSS	2022–2025	Context; validation
Victim categories	SATP / SAIR	2009–2022 (partial)	Context
GTI score and rank	IEP Global Terrorism Index	Editions 2017–2026	Cross-national benchmarking
Digital exposure	DataReportal; NapoleonCat	2024–2026	Online-exposure covariates
CT response	PIPS, PICSS, ISPR, CRSS	2016–2025 (partial)	Intervention context
Legal instruments	Official statutes, UN, EU, CoE	24 instruments, 1966–2025	Governance layer

Note: PIPS = Pak Institute for Peace Studies; PICSS = Pakistan Institute for Conflict and Security Studies; CRSS = Center for Research and Security Studies; SATP = South Asia Terrorism Portal; IEP = Institute for Economics & Peace. Full URLs are listed in the data sources section.

Table 2. Descriptive Statistics, Pakistan 2006–2025 (PIPS)

Variable	N	Mean	SD	Min	Max	Skew.	Ex. kurt.	JB (p)
Attacks	20	977.6	793.4	146	2,586	0.64	-1.06	2.30 (0.316)
Killed	20	1423.4	1014.7	220	3,448	0.60	-1.02	2.05 (0.358)
Lethality	20	1.627	0.386	1.055	2.276	0.60	-0.84	1.77 (0.412)
log Attacks	20	6.517	0.921	4.984	7.858	-0.02	-1.39	1.62 (0.444)
log Killed	20	6.978	0.816	5.394	8.146	-0.25	-1.00	1.04 (0.595)
Δlog Attacks	19	0.003	0.378	-0.657	0.835	0.33	-0.47	0.52 (0.770)
Δlog Killed	19	0.007	0.452	-0.511	1.335	1.26	1.86	7.76 (0.021)

Note: Lethality = killed per attack. JB = Jarque–Bera normality test. Source: Authors' calculations from PIPS data.

Over 2006–2025, PIPS recorded 19,553 attacks and 28,468 deaths. Mean lethality was 1.63 deaths per attack (range 1.06–2.28). Only the differenced fatality series departs from normality (JB $p = 0.021$), reflecting the extreme 2007 jump.

Econometric Core

Integration Properties

We test for unit roots with the augmented Dickey Fuller (ADF) test (Dickey & Fuller, 1979) with lag order selected by AIC. Because asymptotic critical values are unreliable when $T = 20$, we compute finite-sample critical values from MacKinnon's (2010) response surfaces and also a sieve-bootstrap p-value from

999 resamples generated under the unit-root null. We complement this with the KPSS stationarity test (Kwiatkowski et al., 1992) and the Zivot and Andrews (1992) test, which allows an endogenous break in intercept and trend:

$$\Delta y_t = \mu + \beta t + \theta DU_t(T_b) + \gamma DT_t(T_b) + \alpha y_{t-1} + \Sigma_{j=1}^{p} c_j \Delta y_{t-j} + \varepsilon_t \quad (1)$$

Multiple Structural Breaks

Following Bai and Perron (1998, 2003), we estimate a segmented trend model for log attacks, $y_t = a_j + b_j t + u_t$ for t in regime $j = 1,\ldots,m+1$, with all coefficients allowed to change. The global minimizer of the sum of squared residuals for each m is found by dynamic programming with a minimum segment length $h = 4$ years, and the number of breaks is selected by BIC and the modified Schwarz criterion of Liu, Wu and Zidek (LWZ).

Markov-Switching Regimes

To obtain real-time regime probabilities, we estimate a two-state Markov-switching model (Hamilton, 1989) for the growth rate of attacks, $\Delta y_t = \mu_{St} + \sigma_{St}\varepsilon_t$, where $S_t \in \{1, 2\}$ follows a first-order Markov chain with transition probabilities p_{ij}. The Hamilton filter gives filtered probabilities $\Pr(S_t = j \mid I_t)$, which use only information available at time t and are therefore genuine real-time warning signals; the Kim smoother gives full-sample probabilities. Parameters are estimated by maximum likelihood with 30 random starts; standard errors come from the numerical Hessian.

Negative-Binomial Log-Linear INGARCH

Attack counts are overdispersed and persistent. We therefore estimate a log-linear Poisson-autoregression of the type developed by Fokianos et al. (2009) and Fokianos and Tjøstheim (2011), with a negative-binomial (NB2) conditional distribution and regime covariates:

$$Y_t \mid F_{t-1} \sim \mathrm{NB}(\lambda_t, \varphi), \quad \nu_t = \log \lambda_t = d + a_1 \nu_{t-1} + b_1 \log(1 + Y_{t-1}) + \gamma_1 \mathrm{NAP}_t + \gamma_2 \mathrm{KABUL}_t \quad (2)$$

where $\mathrm{NAP}_t = 1$ for $t \geq 2015$ (first full year after the National Action Plan and Operation Zarb-e-Azb) and $\mathrm{KABUL}_t = 1$ for $t \geq 2022$ (first full year after the Taliban takeover of Kabul in August 2021). The NB2 variance is $\lambda_t + \lambda_t^2/\varphi$. Stationarity requires $|a_1 + b_1| < 1$, which we impose in estimation. Inference uses the inverse numerical Hessian; adequacy is checked with a Ljung–Box test on Pearson residuals.

Lethality Elasticity and Cointegration

We estimate $\log K_t = \alpha + \beta \log A_t + \delta_1 \mathrm{NAP}_t + \delta_2 \mathrm{KABUL}_t + \tau t + e_t$ with Newey–West (1987) HAC standard errors, where β is the elasticity of fatalities (K) to attacks (A); $\beta = 1$ means proportional lethality. Long-run proportionality is examined with the Johansen (1991) trace and maximum-eigenvalue tests in a VECM with one lagged difference and an unrestricted constant, with the Reinsel–Ahn small-sample correction, and with the Engle–Granger residual test.

VAR Dynamics

A VAR(1) in $\Delta\log A$ and $\Delta\log K$ is used for Granger causality, generalised impulse responses (Pesaran & Shin, 1998) with 90% residual-bootstrap bands (1,000 replications) and Cholesky forecast-error variance decompositions.

Latent Threat Level and Momentum

We decompose log attacks with a local linear trend model (Harvey, 1989): $y_t = \mu_t + \varepsilon_t$, $\mu_{t+1} = \mu_t + \beta_t + \eta_t$, $\beta_{t+1} = \beta_t + \zeta_t$. The Kalman filter yields the one-sided (real-time) estimate of the slope β_t interpreted as *threat momentum* and the Rauch–Tung–Striebel smoother yields the latent level. Variances are estimated by exact maximum likelihood with diffuse initialization.

Critical-Slowing-Down Indicators

If the 2020 trough was a critical transition, residual variance and lag-1 autocorrelation should rise before it (Dakos et al., 2012; Scheffer et al., 2009). We Gaussian-detrend log attacks for 2006–2020, compute both indicators in rolling seven-year windows, summarize their trends with Kendall's τ, and assess significance against 2,000 AR(1) surrogate series.

Deep Learning Module

Because no deep learning library was available in the computing environment, we implemented an LSTM network from first principles in NumPy, with analytic back-propagation through time, Adam optimization (Kingma & Ba, 2015), gradient clipping and L2 regularization. The analytic gradients were verified against central finite differences (relative error 8.8e-9). Each LSTM has eight hidden units and reads a three-year sequence of (log attacks, log killed). The deep multilayer perceptron (MLP) has three hidden layers (32–16–8, tanh) trained by L-BFGS with weight decay. To reduce variance, each architecture is an ensemble over random seeds (12 LSTMs, 20 MLPs).

Two variants are estimated. The *level* variant predicts log attacks directly. The *residual-learning* (Δ) variant predicts the change in log attacks from lagged changes, lagged change in log killed and lagged lethality, and adds the prediction to the last observed level; this is the standard remedy for non-stationary inputs. Hybrid ensembles average the MLP and LSTM, and a further combination adds damped trend smoothing.

All models are evaluated by an expanding-window, rolling-origin design: for each test year from 2016 to 2025, every model including scaling constants and hyperparameters is re-estimated on data available up to the previous year only. Baselines are a random walk, an ARIMA(1,1,0), damped Holt smoothing, and the NB-INGARCH model. Accuracy is measured by RMSE of log forecasts, mean absolute error in attack counts, MAPE, Theil's *U* relative to the random walk, and directional accuracy. Equal predictive accuracy against the random walk is tested with the Diebold and Mariano (1995) statistic with the Harvey–Leybourne–Newbold small-sample correction.

Explainability Module

For the MLP ensembles we compute *exact* Shapley values by enumerating all $2^5 = 32$ feature coalitions, with interventional expectations taken over the training data:

$$\varphi_i = \Sigma_{S \subseteq N \setminus \{i\}} \, [|S|!(|N| - |S| - 1)!/|N|!] \, [v(S \cup \{i\}) - v(S)] \qquad (3)$$

Exact enumeration guarantees the efficiency property (attributions sum to the prediction minus the baseline); the numerical additivity error was 5.6e-17. For the LSTM ensembles we compute integrated gradients over the input sequence with a zero (training-mean) baseline and 64 Riemann steps

(Sundararajan et al., 2017); the completeness error was 0.017. Permutation importance (30 permutations per feature) and a counterfactual analysis complete the module.

Graph-Based Radicalization-Trajectory Module

Actor-to-province attack counts are not published, but both margins are: in 2024, PIPS attributes 335 attacks to religiously inspired militants, 175 to nationalist insurgents and 11 to sectarian groups, and distributes the same 521 attacks across Khyber Pakhtunkhwa (295), Balochistan (202), Sindh (12), Punjab (11) and Islamabad (1). We recover the latent bipartite edge matrix **E** by iterative proportional fitting (IPF), the maximum-entropy solution consistent with both margins and a prior affinity matrix that encodes documented operational theatres (core = 1, secondary = 0.25, residual = 0.02). Sensitivity to the prior is assessed by 500 random rescaling's of all non-core weights by factors between 0.5 and 2.

On the resulting heterogeneous graph (actor classes, provinces and the named groups TTP and BLA) we compute weighted PageRank, eigenvector and betweenness centrality. A simplified graph convolution (Wu et al., 2019) propagates standardized node features (log attacks, log killed, lethality) over the normalized adjacency $\mathbf{S} = \tilde{\mathbf{D}}^{-1/2}(\mathbf{A} + \mathbf{I})\tilde{\mathbf{D}}^{-1/2}$ for $K = 2$ hops, $\mathbf{H} = \mathbf{S}^2\mathbf{X}$, and a ridge regression maps **H** to provincial attack growth in 2025, evaluated by leave-one-province-out cross-validation. Edge-occlusion attribution, in the spirit of GNNExplainer (Ying et al., 2019), removes each edge in turn and records the change in each province's prediction. Finally, actor growth trajectories are estimated by quasi-Poisson log-linear regression and pooled with a DerSimonian–Laird random-effects model, giving empirical-Bayes growth estimates; compositional change in the actor mix is measured by the Aitchison distance.

Robustness Design

Three checks address data quality and small-sample fragility: (i) a Monte Carlo experiment in which the eight unverified years are perturbed by independent uniform shocks of ±15% (200 replications), re-estimating the INGARCH model, the Bai–Perron breaks and the lethality elasticity each time; (ii) re-estimating the elasticity on verified years only; and (iii) comparing deep models in level and residual form.

Results

Integration Properties and Structural Change

Table 3. Unit-Root and Stationarity Tests

Series	ADF τ (lag)	5% CV	Bootstrap p	KPSS (level)	KPSS (trend)	ZA stat (break)
log Attacks	-0.81 (1)	-3.65	0.940	0.507**	0.113	-4.23 (2019)
log Killed	-1.92 (2)	-3.65	0.551	0.506**	0.106	-3.20 (2018)
Lethality	-3.11 (1)	-3.65	0.125	0.250	0.084	-4.58 (2018)
Δlog Attacks	-2.92 (0)*	-3.03	0.057	0.187	—	—
Δlog Killed	-5.50 (0)***	-3.03	0.004	0.169	—	—

Note: Levels tested with constant and trend; differences with constant. Finite-sample CVs from MacKinnon (2010). Bootstrap p from 999 sieve-bootstrap replications under H0. KPSS 5% CVs: 0.463

(level), 0.146 (trend). Zivot–Andrews model C, 5% CV −5.08; the year reported is the last year of the pre-break regime. *, , * denote rejection at 1%, 5%, 10%.

Table 3 indicates that log attacks and log fatalities are integrated of order one: ADF tests do not reject a unit root (bootstrap $p = 0.94$ and 0.55), the KPSS level test rejects stationarity at 5%, and the differenced series are stationary (Δlog attacks at the 10% level, bootstrap $p = 0.057$; Δlog killed at 1%). The Zivot–Andrews test does not reject the unit root even when a break is allowed, but it places the most likely break at 2019/2020 for attacks and 2018/2019 for fatalities. Lethality behaves differently: the KPSS test does not reject stationarity at either specification, so deaths per attack fluctuate around a stable mean, which anticipates the cointegration result below.

Table 4. Bai–Perron Multiple Structural Breaks in the Trend of Log Attacks

Breaks (m)	Break years (first year of new regime)	SSR	BIC	LWZ
0	—	7.462	-13.73	-0.581
1	2015	3.657	-19.00	-0.663
2	2010, 2020	0.489	-50.27	-2.003
3	2010, 2015, 2020	0.194	-59.79	-2.192

Note: Model $y_t = a_j + b_j t + u_t$, all coefficients break; minimum segment 4 years; global optimum by dynamic programming. sup-F(1|0) = 7.80. For log killed, BIC selects m = 2 with breaks at 2010 and 2020.

Both information criteria select three breaks, with new regimes starting in 2010, 2015 and 2020 (Table 4). The four regimes have a clear interpretation: a peak-escalation phase (2006–2009); a high-plateau decline (2010–2014) following military operations in Swat and South Waziristan; a steep decline (2015–2019) coinciding with the National Action Plan and Operation Zarb-e-Azb; and a renewed escalation from 2020. Fatalities share the 2010 and 2020 breaks. This supports **H1**. Figure 2 overlays these breaks with the Markov-switching regimes.

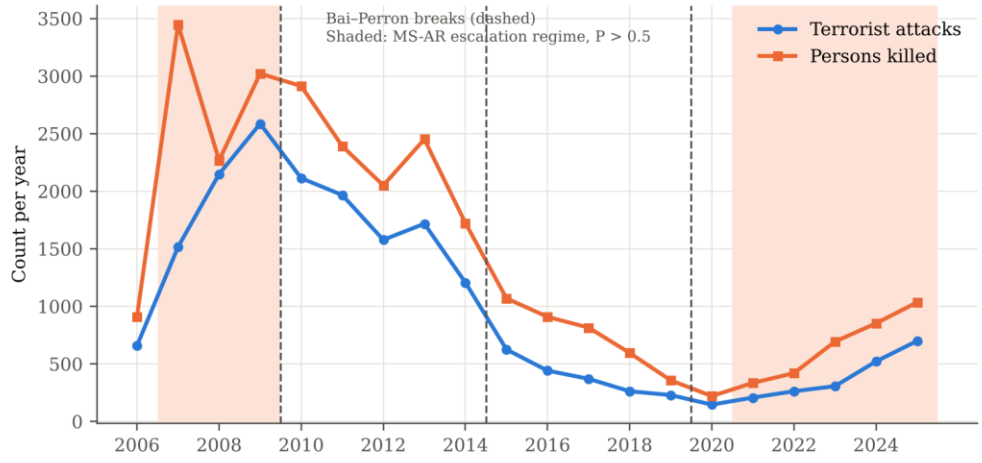


Figure 2. Terrorist attacks and deaths in Pakistan, 2006–2025, with Bai–Perron break dates (dashed lines; first year of new regime) and Markov-switching escalation regime (shaded where the smoothed probability exceeds 0.5). Source: PIPS data; authors' estimates.

Regime Switching: Real-Time Detection of the 2021 Turning Point

Table 5. Two-State Markov-Switching Model for Δlog Attacks

Parameter	Escalation regime (S = 1)	De-escalation regime (S = 2)
Mean growth μ	0.366*** (0.078)	-0.258*** (0.061)
Volatility σ	0.213	0.195
Staying probability p_{jj}	0.906	0.873
Expected duration (years)	10.6	7.8
Log-likelihood (MS / linear)	-2.94 / -7.98	
AIC (MS / linear)	17.89 / 19.96	
LR statistic (non-standard)	10.07	

Note: Standard errors (numerical Hessian) in parentheses. Because the switching parameters are not identified under the linear null, the LR statistic has a non-standard distribution (Hansen, 1992); we therefore rely on AIC and the economic interpretability of the regimes.

The escalation regime implies average growth of 36.6% per year in log terms and the de-escalation regime -25.8%; both regimes are highly persistent, with expected durations of 10.6 and 7.8 years. The switching model is preferred by AIC (17.89 vs. 19.96).

The key early-warning result concerns the *filtered* probabilities, which use only data available at the time. The filtered probability of the escalation regime was no higher than 0.01 in 2018, 2019 and 2020, and jumped to **0.94** in 2021 and 0.99 in 2022. In other words, the filter would have signalled a shift into a persistent escalation regime at the end of 2021, when PIPS recorded only 207 attacks—less than one-third of the 2025 level and well below any historical alarm threshold based on levels. This is the type of anticipation that a level-based alert would miss.

Count Dynamics and Policy Regimes: NB-INGARCH

Table 6. Negative-Binomial Log-Linear INGARCH(1,1) Model for Annual Attacks

Parameter	Estimate	Std. error	z	p-value
d (intercept)	3.388***	0.750	4.52	<0.001
a1 (persistence ν_{t-1})	-0.088	0.219	-0.40	0.689
b1 (feedback log(1+y_{t-1}))	0.645***	0.223	2.90	0.004
γ1 NAP/Zarb-e-Azb regime (≥2015)	-0.982***	0.181	-5.43	<0.001
γ2 post-Kabul regime (≥2022)	0.453***	0.157	2.89	0.004
log φ (NB dispersion)	3.110***	0.337	9.22	<0.001

Note: Log-likelihood = -121.87; AIC = 255.74 (model without regime covariates: 270.11). LR test of $\gamma_1 = \gamma_2 = 0$: $\chi^2(2) = 18.38$, p = 0.0001. Ljung–Box Q(3) on Pearson residuals = 2.05 (p = 0.562). NB dispersion φ = 22.4. ***,** , * = 1%, 5%, 10%.

The INGARCH estimates (Table 6) support **H2**. Conditional on count persistence, the NAP/Zarb-e-Azb

regime is associated with a 62.6% lower expected attack intensity (exp(-0.982) − 1), and the post-Kabul regime with a 57.2% higher intensity; both effects are significant at 1%, and jointly they improve fit substantially (LR $\chi^2(2) = 18.38$). Persistence operates through the observation-feedback term ($b_1 = 0.645$), not through the latent intensity (a_1 is small and insignificant), so last year's attacks carry most of the memory. Pearson residuals show no remaining autocorrelation. These are associations with regime timing, not causal effects: the dummies capture everything that changed in those years, including operations, cross-border conditions and group reorganization.

Lethality: Elasticity and Cointegration

Table 7. Fatality–Attack Elasticity (HAC) and Cointegration Tests

Panel A: log Killed regression	**Coefficient**	**HAC s.e.**	**t**	**p**
Constant	0.817	0.657	1.24	0.233
log Attacks	0.946***	0.091	10.37	<0.001
NAP regime (≥2015)	0.394*	0.219	1.79	0.093
Post-Kabul (≥2022)	0.100	0.116	0.86	0.405
Trend	-0.025	0.016	-1.57	0.137
R^2 / N	0.953			20
H0: elasticity = 1 (t)	-0.59			
Panel B: cointegration	Statistic	5% CV		Decision
Johansen trace, r = 0	15.25	15.41		Not rejected (marginal)
Johansen trace, r ≤ 1	3.39	3.76		Not rejected
Johansen max-eig, r = 0	11.86	14.07		Not rejected
Trace r = 0, Reinsel–Ahn corrected	11.86	15.41		Not rejected
Engle–Granger residual ADF	-4.21	-3.67		Rejected: cointegration

Note: Newey–West bandwidth 2. Johansen VECM with one lagged difference, unrestricted constant; Osterwald-Lenum critical values. Engle–Granger cointegrating slope = 0.861; approximate small-sample 5% CV.

The elasticity of fatalities with respect to attacks is 0.946 (HAC s.e. 0.091) and is statistically indistinguishable from one (t = -0.59), so in the long run deaths scale proportionally with attacks. The post-2015 dummy is positive (0.394, $p = 0.093$), consistent with the rise in deaths per attack in 2016–2018 as fewer but more lethal attacks were carried out. Cointegration evidence is mixed: The Johansen trace statistic (15.25) falls just short of the 5% critical value (15.41), while the single-equation Engle–Granger test rejects no-cointegration. With twenty observations the Johansen test has low power, so we read the evidence as moderately supportive of **H3**: lethality is stable around a long-run mean and attack counts are the primary driver of fatalities.

VAR Dynamics

The VAR(1) in growth rates is stable (largest root 0.428). Neither Granger-causality test is significant at conventional levels (Δlog killed → Δlog attacks: $p = 0.576$; Δlog attacks → Δlog killed: $p = 0.235$). However, the variance decomposition is asymmetric: at a five-year horizon, 59.5% of the forecast-error variance of fatality growth is explained by attack shocks, while fatality shocks explain only 0.9% of attack growth. Generalised impulse responses (Figure 3) show that both shocks die out within three years; attack shocks transmit contemporaneously to fatalities. For early warning, this means that monitoring attack frequency is sufficient to track fatality risk, but fatalities add short-run information about the severity of the current regime.

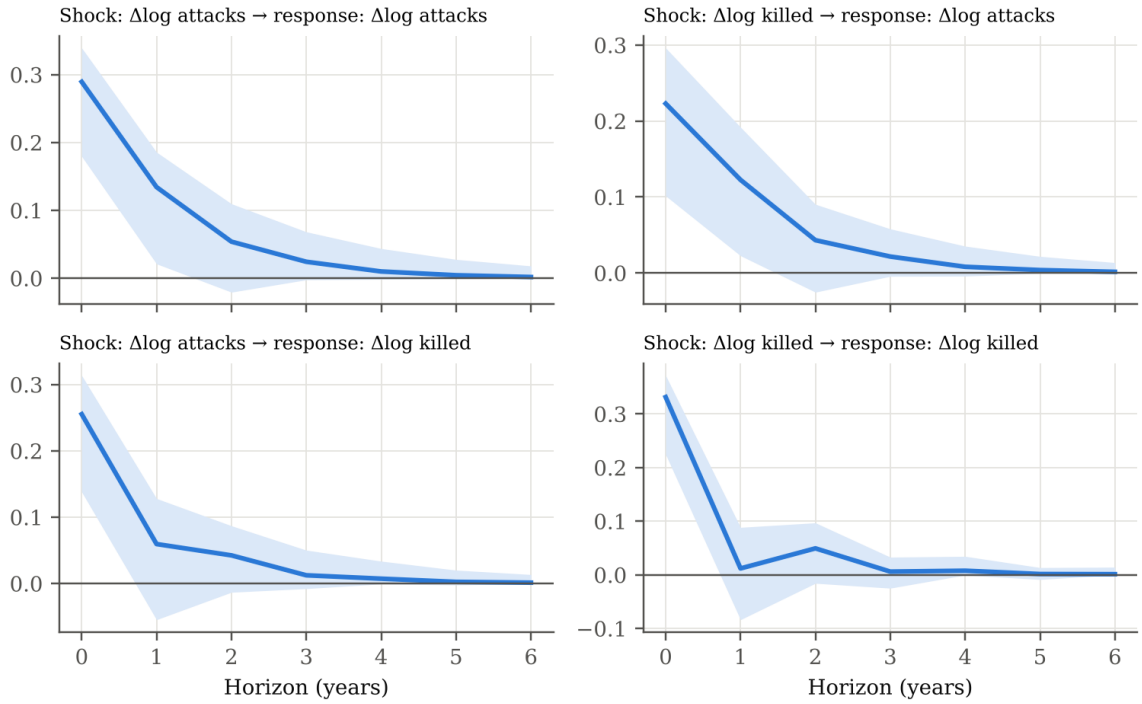


Figure 3. Generalized impulse responses from the VAR(1) in Δlog attacks and Δlog killed, with 90% residual-bootstrap bands (1,000 replications). Source: Authors' estimates.

Latent Threat Momentum and Critical-Slowing-Down Signals

The unobserved-components model attributes almost all trend variation to the slope ($\sigma_\zeta = 0.237$; $\sigma_\eta \approx 0$), producing a smooth latent level (Figure 4a). The one-sided slope estimate threat momentum was negative in every year from 2010 to 2020 except a negligible +0.01 in 2013 (−0.37 in 2020), turned positive in 2021 (+0.15) and reached +0.44 in 2024, the first year in which the 90% band lies entirely above zero (Figure 4b). Momentum therefore confirms, with a delay, what the Markov-switching filter signalled in 2021.

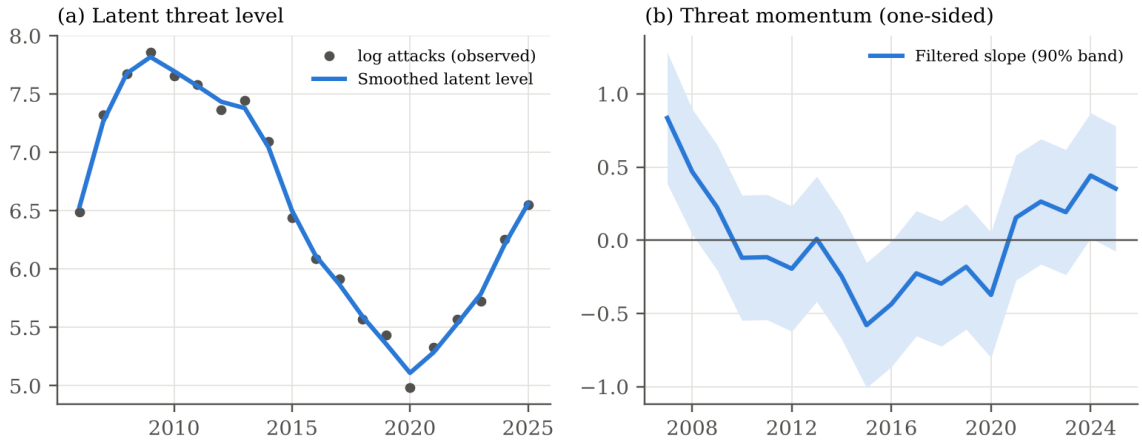


Figure 4. Unobserved-components local linear trend model for log attacks: (a) smoothed latent level; (b) one-sided Kalman-filtered slope (threat momentum) with 90% band. Source: Authors' estimates.

By contrast, classic critical-slowing-down indicators did not provide a statistically reliable warning before the 2020 turning point (Figure 5). Rolling variance and lag-1 autocorrelation both trend upward after 2014 (Kendall $\tau = 0.39$ in each case), but the trends are not significant against AR(1) surrogates ($p = 0.28$ and 0.23). With annual data and only nine windows, generic resilience indicators are too noisy to be used as alarms; regime-specific models are more informative.

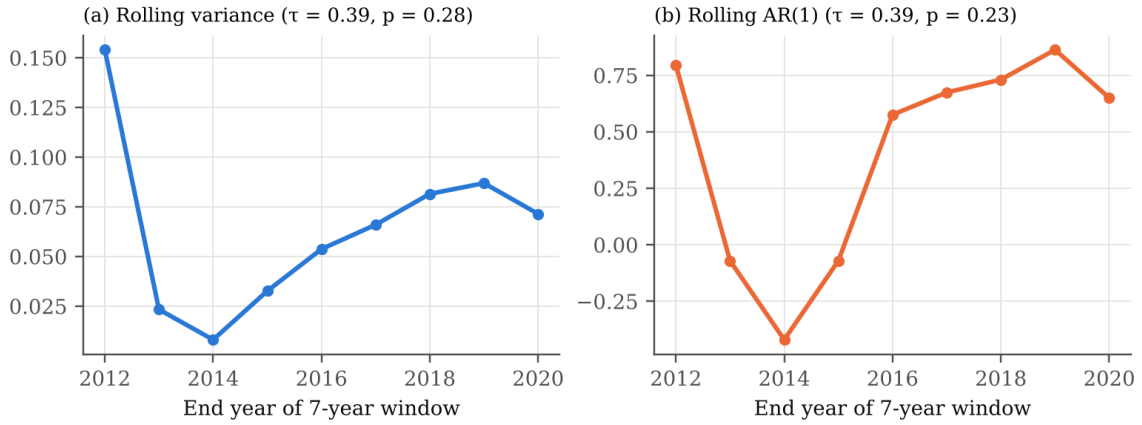


Figure 5. Critical-slowing-down indicators for detrended log attacks, rolling seven-year windows ending 2012–2020; significance from 2,000 AR(1) surrogates. Source: Authors' estimates.

Deep Learning versus Parsimonious Baselines

Table 8. Rolling-Origin One-Step-ahead Forecast Accuracy, 2016–2025 (10 Forecasts)

Model	RMSE (log)	MAE (attacks)	MAPE (%)	Theil's U	Direction acc.	DM vs RW (p)
Random walk	0.326	103.2	30.5	1.000	—	—
AR(1) on Δ (ARIMA(1,1,0))	0.323	72.3	22.6	0.991	0.8	-0.04 (0.97)

Model	RMSE (log)	MAE (attacks)	MAPE (%)	Theil's U	Direction acc.	DM vs RW (p)
Damped Holt	0.283	57.4	19.7	0.869	0.8	-0.78 (0.45)
NB-INGARCH(1,1)	0.318	98.3	30.1	0.974	0.7	-0.38 (0.71)
Deep MLP ensemble	0.525	133.4	36.2	1.609	0.3	1.64 (0.13)
LSTM ensemble	0.382	99.0	30.8	1.170	0.5	1.22 (0.25)
Hybrid DL ensemble	0.424	107.2	30.4	1.300	0.4	1.53 (0.16)
Deep MLP-Δ ensemble	0.478	100.7	33.8	1.465	0.7	1.49 (0.17)
LSTM-Δ ensemble	0.420	87.9	29.1	1.287	0.5	1.22 (0.25)
Hybrid DL-Δ ensemble	0.433	94.7	31.5	1.328	0.6	1.30 (0.23)
Hybrid DL-Δ + Damped Holt	0.375	83.5	28.0	1.149	0.7	0.69 (0.51)

Note: All models re-estimated at each origin using information up to t − 1 only. Theil's U < 1 means better than the random walk. DM = Diebold–Mariano statistic with Harvey–Leybourne–Newbold correction; positive values mean higher squared error than the random walk. Δ = residual (first-difference) learning variant.

Table 8 and Figure 6 report the out-of-sample comparison. The best model is damped Holt smoothing (RMSE 0.283, MAE 57.4 attacks, Theil's *U* = 0.869, direction correct in 8 of 10 years). The NB-INGARCH model and ARIMA(1,1,0) are close to the random walk. All deep learning variants have higher RMSE than the random walk (Theil's *U* between 1.15 and 1.61). Residual learning helps the MLP's directional accuracy (from 0.3 to 0.7) and the LSTM's MAE (from 99.0 to 87.9 attacks), and combining the Δ-ensemble with damped Holt lowers MAPE to 28.0%. Yet no model differs significantly from the random walk (all DM $p > 0.10$). **H4 is therefore not supported.**

Two features of the errors are instructive. First, every model under-predicted the 2024 surge (actual 521 attacks), because the 2023 increase was modest; the largest errors are at turning points, exactly where early warning matters most. Second, the level-based deep models were biased toward the 2009–2013 high-violence era on which they were partly trained and therefore mean-reverted toward it—an illustration of how deep networks extrapolate poorly from short, non-stationary series. As an escalation classifier, the combined Δ-ensemble produced no false alarms (precision 2/2) but detected only 2 of 5 escalation years (2024 and 2025), missing the onset years 2021–2023 that the Markov-switching filter identified.

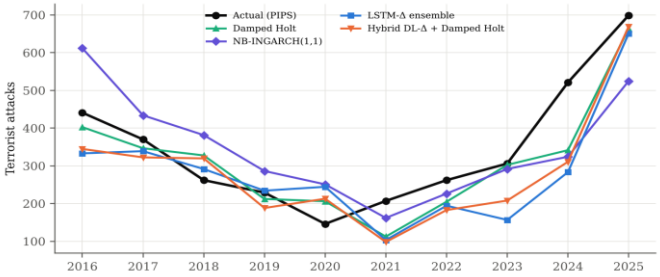


Figure 6. Rolling-origin one-step-ahead forecasts of annual terrorist attacks, 2016–2025, for selected models against PIPS actuals. Source: Authors' estimates.

Explainability: What Drives the Algorithmic Warnings?

Figure 7a reports mean absolute Shapley values for the residual-learning MLP across the ten forecast origins. The most influential inputs are the lagged change in fatalities (0.122) and lagged lethality (0.112), followed by the first and second lags of attack growth (0.068 and 0.062). For the 2025 forecast, the baseline expectation of Δlog attacks was -0.107; the 2024 surge in attacks added +0.254 and the rise in deaths added +0.105, yielding a predicted increase of 0.286 log points (≈ 694 attacks against an actual 699). Integrated gradients for the LSTM-Δ ensemble (Figure 7b) give the same qualitative message: fatality dynamics at $t-1$ and $t-3$ carry the largest attributions.

Permutation importance offers a necessary caution: permuting any single feature slightly *reduced* out-of-sample error, meaning that no individual input is robustly predictive across origins. The explanations are faithful to the model, but the model is fragile. A counterfactual test makes the point concrete: holding 2025 fatalities fixed, the MLP-Δ component predicted further escalation in 2026 for every 2025 attack level between 400 and 1,000 (minimum predicted Δ = 0.080). The warning is therefore driven by multi-year momentum and lethality rather than by any single year's count useful for transparency, and a reason why such outputs must not be presented as precise targets.

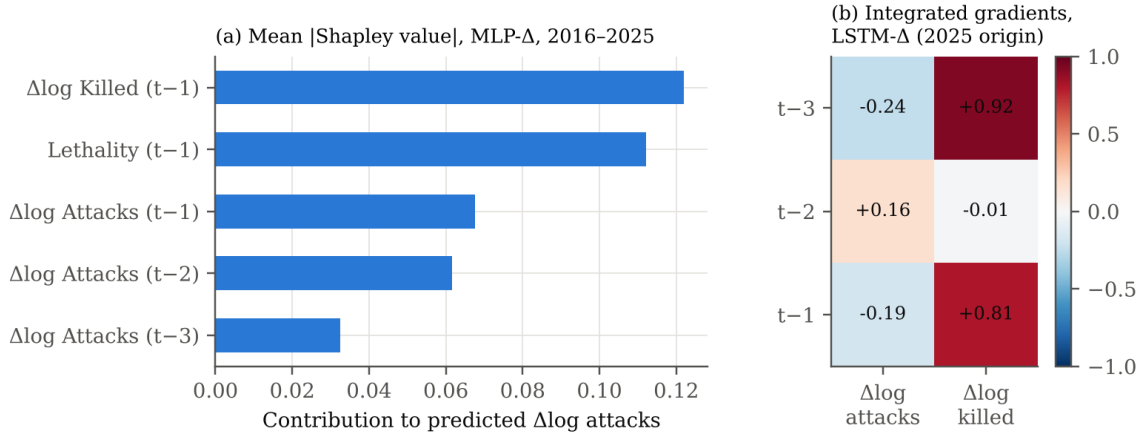


Figure 7. Explainability results: (a) mean absolute exact Shapley values for the residual-learning deep MLP ensemble, 2016–2025 origins; (b) integrated-gradient attributions for the LSTM-Δ ensemble at the 2025 origin (standardized units). Source: Authors' estimates.

Graph-Based Radicalization Trajectories

Table 9. Estimated Actor–Province Attack Edges, 2024 (IPF Maximum-Entropy Reconstruction)

Actor class	KP	Balochistan	Sindh	Punjab	Islamabad	Total
Religiously inspired militants	279 [272–282]	37 [33–47]	9 [7–10]	9 [6–10]	0.9	335
Nationalist insurgents	6 [4–14]	164 [154–167]	3 [2–5]	2 [1–5]	0.0	175
Sectarian	10 [8–10]	1 [1–3]	0 [0–0]	0 [0–0]	0.0	11
Total	295	202	12	11	1	521

Note: Point estimates with 90% ranges across 500 perturbations of non-core prior weights (×0.5 to ×2). Margins from PIPS PSR 2024. KP = Khyber Pakhtunkhwa.

The reconstructed graph (Table 9; Figure 8) shows a clear territorial division: an estimated 279 of 295 attacks in Khyber Pakhtunkhwa are attributable to religiously inspired militants and 164 of 202 attacks in Balochistan to nationalist insurgents, with narrow ranges under prior perturbation. Religiously inspired militancy also projects into Balochistan (≈ 37 attacks), which is the main cross-theatre link. Centrality confirms this structure: religiously inspired militants have the highest PageRank (0.292) and betweenness (0.889), followed by nationalist insurgents (0.173); among provinces, Balochistan has the highest betweenness (0.389) because it is the only province where both major actor classes are strongly present.

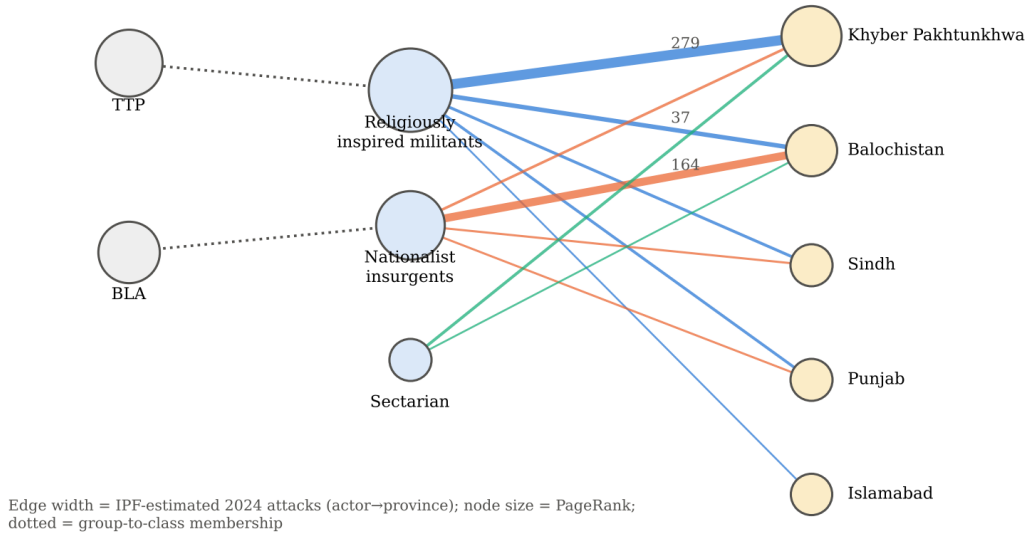


Figure 8. Heterogeneous radicalization graph for 2024: actor classes, provinces and named groups. Edge width is proportional to IPF-estimated attacks; node size to weighted PageRank. Source: Authors' estimates from PIPS and IEP-GTI.

The simplified graph convolution did not improve the prediction of provincial growth in 2025: its attack-weighted leave-one-out RMSE (0.195) was better than the naive mean (0.244) but worse than the same ridge model on unpropagated features (0.148). With five provinces and one transition, graph propagation cannot yet be shown to add predictive value. Edge occlusion is nonetheless informative for audit: the religiously inspired militant–Islamabad edge accounts for the largest single attribution (+0.213 to Islamabad's predicted growth), and the nationalist–Balochistan edge for +0.070 to Balochistan's showing which relational assumptions drive each provincial score.

Table 10. Actor Growth Trajectories: Quasi-Poisson Log-Linear Growth with Empirical-Bayes Pooling

Actor (source, years)	Growth (log/yr)	s.e.	EB estimate	Implied annual growth
Religiously inspired militants (PIPS) (2020–2024)	0.317	0.057	0.320	38%

Actor (source, years)	Growth (log/yr)	s.e.	EB estimate	Implied annual growth
Nationalist insurgents (PIPS) (2020–2024)	0.355	0.130	0.365	44%
Sectarian (PIPS) (2020–2024)	0.123	0.116	0.142	15%
TTP (IEP-GTI) (2023–2025)	0.465	0.185	0.473	61%
Baloch groups incl. BLA (IEP-GTI) (2023–2024)	1.469	0.103	1.437	321%

Note: DerSimonian–Laird pooled growth = 0.547; $\tau^2 = 0.293$; $I^2 = 96.4\%$. The Baloch-groups estimate is based on two IEP-GTI observations and should be read as indicative.

Actor trajectories are highly heterogeneous ($I^2 = 96.4\%$; Table 10, Figure 9). Religiously inspired militancy grew by about 38% a year between 2020 and 2024 and nationalist insurgency by about 44%, while sectarian violence was flat. In the IEP-GTI series, TTP attacks grew by about 61% a year in 2023–2025, and Baloch groups' attacks more than quadrupled between 2023 and 2024. Compositional analysis points the same way: the nationalist share of PIPS-attributed attacks rose from 28.1% in 2023 to 33.6% in 2024, and the Aitchison distance of the 2023–2024 shift (0.580) is more than three times that of 2020–2023 (0.173). **H5 is supported**: the fastest-moving trajectory is nationalist, not religious, which matters for any system that is tuned primarily to religious-extremist signals.

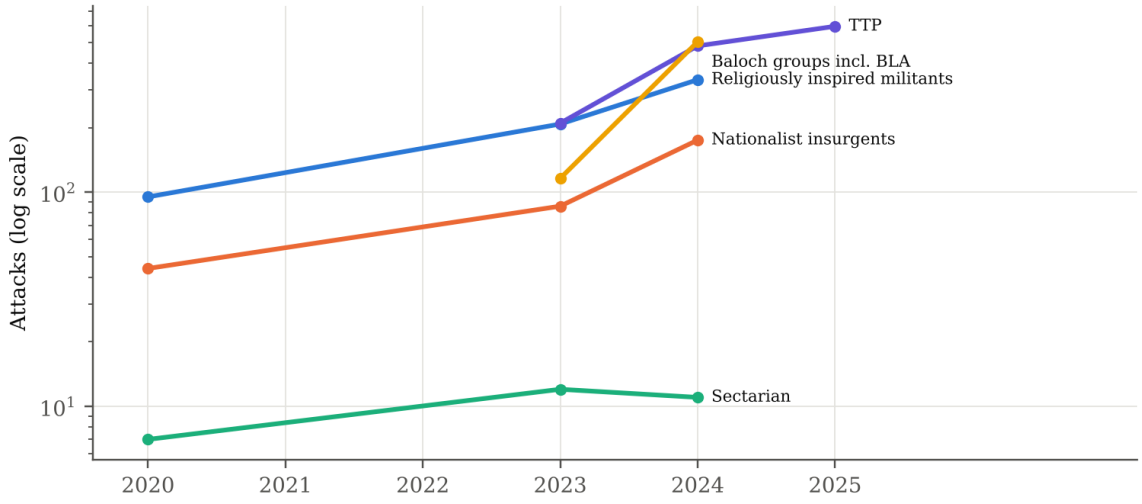


Figure 9. Actor growth trajectories (log scale). PIPS actor classes 2020–2024 and IEP-GTI named groups 2023–2025. Source: PIPS; IEP Global Terrorism Index 2024–2026.

5.10 Robustness

Table 11. Robustness to Unverified years and Estimation Choices

Quantity	Baseline	Perturbation: 5th pct.	Median	95th pct.
INGARCH persistence ($a_1 + b_1$)	0.557	0.493	0.552	0.597
INGARCH post-Kabul effect γ_2	0.453	0.414	0.460	0.494
Lethality elasticity β	0.946	0.864	0.934	0.993
Elasticity, verified years only (n = 12)	0.938 (s.e. 0.078)			
Break selected at 2010 / 2015 / 2020 (share)	Yes	1.00	0.815	0.925

Note: 200 Monte Carlo replications in which the 8 unverified years (2008, 2010, 2011, 2012, 2013, 2014, 2015, 2016) are multiplied by independent U(0.85, 1.15) shocks. In the last row the three numeric columns give the share of replications selecting each break year.

The main results are robust (Table 11; Figure 10a). Under ±15% perturbations of the unverified years, the post-Kabul effect stays between 0.41 and 0.49, persistence between 0.49 and 0.60, and the elasticity between 0.86 and 0.99. The 2010 and 2020 breaks are selected in 100% and 92.5% of replications; the 2015 break in 81.5%, with the remainder shifting it to 2014. On verified years alone the elasticity is 0.938. Figure 10b shows the provincial concentration that underpins the graph module: Khyber Pakhtunkhwa and Balochistan together account for over 95% of attacks in both 2024 and 2025.

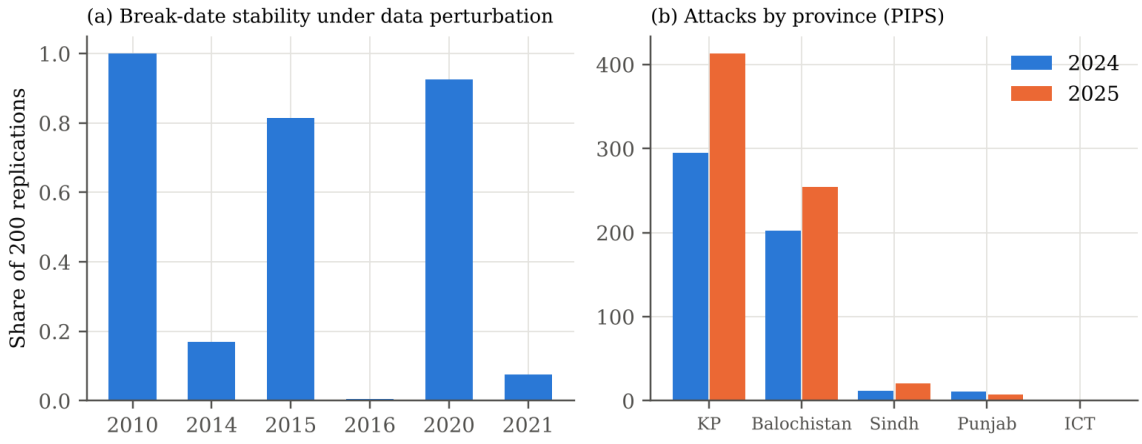


Figure 10. (a) Share of 200 perturbation replications in which each break year is selected; (b) terrorist attacks by province, 2024–2025. Source: PIPS; authors' estimates.

Forecast Outlook for 2026 and Cross-National Position

Trained on all data through 2025, the residual-learning MLP projects 760 attacks for 2026, the LSTM-Δ 996 and damped Holt 883; their combination is 875, with an 80% empirical interval of 579–1,321 derived from out-of-sample errors. The NB-INGARCH model, which embeds the post-Kabul regime, projects 683. All models point to continued escalation; they disagree substantially on its size, and Section 5.7 shows that none has a demonstrated advantage over simple extrapolation. Cross-nationally, Pakistan

moved from second (score 8.374) to first (8.574) in the GTI between the 2025 and 2026 editions, ahead of Burkina Faso (8.324), Niger (7.816) and Nigeria (7.792). The combination of a confirmed escalation regime, positive momentum and a top-ranked GTI position defines the setting in which governance safeguards are most likely to be tested.

The Ethical–Legal Governance Architecture

The results determine the governance design. Early-warning value came from aggregate, regime-level models; deep models added no demonstrated accuracy; graph edges are partly reconstructed rather than observed; and explanations, while faithful, reveal fragile models. A rights-compliant architecture must encode these facts. We propose seven design principles, anchored in the 24 instruments coded in the workbook (Table 12).

- **Aggregate-level purpose limitation.** Outputs are confined to national, provincial and actor-class levels. The framework never produces risk scores for individuals or communities, consistent with the prohibition on profiling-only crime-risk prediction in Article 5 of the EU AI Act and with the ICCPR's protection of privacy, expression and assembly (Arts. 17, 19, 21).

- **Epistemic accountability and benchmark disclosure.** Every deployed model must report its out-of-sample performance against parsimonious baselines with formal tests. Where, as here, deep models do not outperform simple ones, the simpler and more interpretable model is preferred (Rudin, 2019).

- **Human-in-the-loop thresholds.** A regime alert (e.g., filtered escalation probability above 0.8 for one year) triggers analyst review and inter-agency assessment through NACTA, not automated action.

- **Legality and judicial authorization.** Any movement from aggregate warning to individual investigation must follow existing statutory routes the warrant model of the Investigation for Fair Trial Act 2013 and due-process limits on Fourth Schedule listing under the Anti-Terrorism Act 1997 rather than model outputs.

- **Transparency artefacts.** Model cards, data-provenance logs (including the verification status of each data point), data-protection impact assessments and algorithmic impact assessments accompany each version.

- **Contestability and independent oversight.** Explanations (Shapley, integrated gradients, edge occlusion) are retained in an audit trail available to parliamentary and judicial oversight, consistent with the UN Global Counter-Terrorism Strategy's fourth pillar and the Council of Europe Framework Convention.

- **Sunset and re-validation.** Models expire after a fixed period and after any detected structural break, because as the 2010, 2015 and 2020 breaks show the data-generating process changes.

Table 12. Governance Matrix: Framework Modules, Risks and Legal Anchors

Module	Principal rights risk	Safeguard	Legal anchor
Temporal econometric core	Over-reliance on point forecasts	Interval reporting; regime probabilities; benchmark disclosure	OECD AI Principles (2019); CoE Framework Convention (2024)
Deep learning ensembles	Opacity; spurious precision	Exact Shapley, IG; deploy only if superior to baselines	EU AI Act (high-risk law-enforcement uses)
Graph trajectory model	Guilt by association; community stigmatization	Actor-class and province level only; no person nodes	ICCPR Arts. 17, 21; UN PVE Plan of Action (A/70/674)
Online-exposure branch	Chilling effects on expression	Population-level reach statistics only; no content surveillance	ICCPR Art. 19; PECA 2016 s.37; Christchurch Call
CT response / intervention	Endogeneity; feedback loops	Lagged or instrumented use; no automatic targeting	UNSC Res. 2396 (safeguards); ATA 1997
Cross-agency data fusion	Function creep	Purpose limitation; DPIA; access logging	NACTA Act 2013; IFTA 2013
Content-removal interface	Over-removal of lawful speech	Notice, appeal and transparency reporting	EU TCO Regulation 2021/784; PECA (Amendment) 2025

Note: The matrix is a normative design proposal coded by the authors from the instruments in the workbook; it is not an empirical estimate.

This architecture also responds to Pakistan's specific legal trajectory. The 2025 amendment to PECA created a social-media regulatory authority and an offence of disseminating false information, which civil-society groups have criticized for its potential effect on free expression. An early-warning system linked to such powers would amplify that risk. By restricting the online-exposure branch to population-level reach statistics and separating warning from enforcement, the framework keeps algorithmic anticipation from becoming a channel for expanding speech controls.

Discussion

Anticipation works at the level of regimes, not points. The most useful warning in this study came from a two-state Markov-switching filter, which signalled the 2021 regime change in real time, and from the INGARCH model, which tied the escalation to the post-August-2021 environment. Both models are small, interpretable and auditable. This is consistent with Cederman and Weidmann's (2017) argument that conflict forecasting should focus on what is predictable persistence and regime structure rather than precise counts.

Deep learning needs data it does not yet have. With twenty annual observations, the LSTM and MLP ensembles could not learn more than simple trend extrapolation does. This is not a negative verdict on deep learning for counter-extremism; it identifies the condition under which it could add value: high-frequency, sub-national, multimodal event data. The workbook's monthly sheet shows how fragmentary such data currently are in the public domain. Investment in standardized, open, event-level data with the

definitional consistency that PIPS, PICSS, CRSS and IEP currently lack is a precondition for the deep-learning layer, not an afterthought.

Radicalization is plural. The graph module shows two largely separate theatres religiously inspired militancy in Khyber Pakhtunkhwa and nationalist insurgency in Balochistan linked through Balochistan. The fastest growth is in the nationalist theatre. Early-warning systems built around religious-extremist content would therefore be poorly calibrated to the fastest-growing source of violence, and counter-extremism policy that treats all violence as ideologically homogeneous risks misallocating resources and alienating communities with distinct grievances.

Accountability must be epistemic as well as legal. Governance debates typically assume that predictive models work and ask how to constrain them. Our results show that the first accountability question is whether the model works better than a transparent alternative. Requiring benchmark disclosure would protect both rights and effectiveness: it prevents opaque systems from being justified by claims of superior accuracy that the evidence does not support.

Limitations. The core series is short ($T = 20$) and eight years are unverified, although robustness checks show that the main inferences survive plausible errors. Regime dummies capture timing, not causal effects. The actor–province graph is reconstructed from margins and documented theatres and should be replaced by event-level attribution when available. The digital-exposure, counter-terrorism response and monthly sheets are too sparse to enter the predictive models and serve as context. Finally, the deep-learning results depend on this data regime and should be re-tested on richer data.

Conclusion and Policy Implications

This paper developed and tested a framework that moves from algorithmic anticipation to normative accountability in early warning of violent extremism. Using real data for Pakistan, it found that violence moves through persistent regimes with breaks in 2010, 2015 and 2020; that a Markov-switching filter detected the 2021 escalation in real time; that the post-Kabul period is associated with a 57% higher attack intensity; that fatalities scale proportionally with attacks; that deep learning ensembles did not outperform transparent baselines; and that nationalist insurgency is the fastest-growing trajectory. These findings support H1, H2, H3 and H5, and reject H4.

Five policy implications follow. First, national early-warning capacity should be built around regime-aware, interpretable models, with deep learning added only where it demonstrably improves accuracy. Second, public authorities should invest in an open, standardized, sub-national event dataset with consistent definitions. Third, warning systems should be calibrated to both religiously inspired and nationalist violence. Fourth, all predictive outputs should be aggregate, interval-based and subject to human review, with any move to individual measures channelled through judicial authorization. Fifth, model cards, benchmark disclosure, audit trails and sunset clauses should be required by regulation or NACTA guidance. Future research should apply the framework to monthly district-level event data, introduce causal identification for counter-terrorism interventions, and test the governance architecture through structured evaluation with oversight bodies.

References

1. Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica, 66*(1), 47–78. https://doi.org/10.2307/2998540
2. Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models.

Journal of Applied Econometrics, 18(1), 1–22. https://doi.org/10.1002/jae.659
3. Borum, R. (2011). Radicalization into violent extremism I: A review of social science theories. *Journal of Strategic Security, 4*(4), 7–36. https://doi.org/10.5038/1944-0472.4.4.1
4. Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review, 82*(5), 977–1008. https://doi.org/10.1177/0003122417725865
5. Cederman, L.-E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science, 355*(6324), 474–476. https://doi.org/10.1126/science.aal4483
6. Clauset, A., Young, M., & Gleditsch, K. S. (2007). On the frequency of severe terrorist events. *Journal of Conflict Resolution, 51*(1), 58–87. https://doi.org/10.1177/0022002706296157
7. Conway, M. (2017). Determining the role of the internet in violent extremism and terrorism: Six suggestions for progressing research. *Studies in Conflict & Terrorism, 40*(1), 77–98. https://doi.org/10.1080/1057610X.2016.1157408
8. Council of Europe. (2024). *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* (CETS No. 225). Council of Europe.
9. Dakos, V., Carpenter, S. R., Brock, W. A., Ellison, A. M., Guttal, V., Ives, A. R., Kéfi, S., Livina, V., Seekell, D. A., van Nes, E. H., & Scheffer, M. (2012). Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS ONE, 7*(7), e41010. https://doi.org/10.1371/journal.pone.0041010
10. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431. https://doi.org/10.2307/2286348
11. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599
12. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608
13. Enders, W., & Sandler, T. (2005). After 9/11: Is it all different now? *Journal of Conflict Resolution, 49*(2), 259–277. https://doi.org/10.1177/0022002704272833
14. Enders, W., Sandler, T., & Gaibulloev, K. (2011). Domestic versus transnational terrorism: Data, decomposition, and dynamics. *Journal of Peace Research, 48*(3), 319–337. https://doi.org/10.1177/0022343311398926
15. Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica, 55*(2), 251–276. https://doi.org/10.2307/1913236
16. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research, 81*, 160–171.
17. European Parliament & Council of the European Union. (2021). Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online. *Official Journal of the European Union, L 172*, 79–109.
18. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union, L series*.
19. Fair, C. C. (2014). *Fighting to the end: The Pakistan army's way of war*. Oxford University Press.
20. Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.
21. Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., & Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. In E. Spiro & Y.-Y. Ahn (Eds.), *Social Informatics: SocInfo 2016* (pp. 22–39). Springer. https://doi.org/10.1007/978-3-319-47874-6_3

22. Fokianos, K., & Tjøstheim, D. (2011). Log-linear Poisson autoregression. *Journal of Multivariate Analysis, 102*(3), 563–578. https://doi.org/10.1016/j.jmva.2010.11.002
23. Fokianos, K., Rahbek, A., & Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association, 104*(488), 1430–1439. https://doi.org/10.1198/jasa.2009.tm08270
24. Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes. *Criminology & Public Policy, 16*(1), 99–117. https://doi.org/10.1111/1745-9133.12249
25. Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica, 57*(2), 357–384. https://doi.org/10.2307/1912559
26. Hansen, B. E. (1992). The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics, 7*(S1), S61–S82. https://doi.org/10.1002/jae.3950070506
27. Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.
28. Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., Hoyles, F., Hultman, L., Högbladh, S., Jansen, R., Mouhleb, N., Muhammad, S. A., Nilsson, D., Nygård, H. M., Olafsdottir, G., Petrova, K., Randahl, D., Rød, E. G., Schneider, G., … Vestby, J. (2019). ViEWS: A political violence early-warning system. *Journal of Peace Research, 56*(2), 155–174. https://doi.org/10.1177/0022343319823860
29. Hegre, H., Bell, C., Colaresi, M., Croicu, M., Hoyles, F., Jansen, R., Leis, M. R., Lindqvist-McGowan, A., Randahl, D., Rød, E. G., & Vesco, P. (2021). ViEWS2020: Revising and evaluating the ViEWS political violence early-warning system. *Journal of Peace Research, 58*(3), 599–611. https://doi.org/10.1177/0022343320962157
30. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
31. Institute for Economics & Peace. (2024). *Global Terrorism Index 2024: Measuring the impact of terrorism*. IEP. https://www.economicsandpeace.org/wp-content/uploads/2024/02/GTI-2024-web-290224.pdf
32. Institute for Economics & Peace. (2025). *Global Terrorism Index 2025*. IEP.
33. Institute for Economics & Peace. (2026). *Global Terrorism Index 2026*. IEP.
34. Jadoon, A. (2021). *The Islamic State in Afghanistan and Pakistan: Strategic alliances and rivalries*. United States Institute of Peace.
35. Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica, 59*(6), 1551–1580. https://doi.org/10.2307/2938278
36. Johnson, N. F., Zheng, M., Vorobyeva, Y., Gabriel, A., Qi, H., Velasquez, N., Manrique, P., Johnson, D., Restrepo, E., Song, C., & Wuchty, S. (2016). New online ecology of adversarial aggregates: ISIS and beyond. *Science, 352*(6292), 1459–1463. https://doi.org/10.1126/science.aaf0675
37. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1412.6980
38. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1609.02907
39. Kundnani, A., & Hayes, B. (2018). *The globalization of countering violent extremism policies: Undermining human rights, instrumentalising civil society*. Transnational Institute.

40. Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y
41. Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x
42. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
43. MacKinnon, J. G. (2010). *Critical values for cointegration tests* (Queen's Economics Department Working Paper No. 1227). Queen's University.
44. McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence, 20*(3), 415–433. https://doi.org/10.1080/09546550802073367
45. Moghaddam, F. M. (2005). The staircase to terrorism: A psychological exploration. *American Psychologist, 60*(2), 161–169. https://doi.org/10.1037/0003-066X.60.2.161
46. Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association, 106*(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546
47. Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association, 110*(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710
48. Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review, 112*(2), 358–375. https://doi.org/10.1017/S0003055417000570
49. Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708. https://doi.org/10.2307/1913610
50. OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). OECD.
51. Pak Institute for Peace Studies. (2025). *Pakistan Security Report 2024*. PIPS. https://pakpips.com/web/wp-content/uploads/2025/01/Overview_PIPS-Security-Report-2024.pdf
52. Pak Institute for Peace Studies. (2026). *Pakistan Security Report 2025*. PIPS. https://www.pakpips.com/?p=16138
53. Pakistan Institute for Conflict and Security Studies. (2026, January 1). *Pakistan's 2025 counterterrorism toll surges 74% as militancy hits multi-year highs* [Press release]. PICSS. https://www.picss.net/
54. Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters, 58*(1), 17–29. https://doi.org/10.1016/S0165-1765(97)00214-0
55. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
56. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, 94*, 15–55.
57. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

https://doi.org/10.1038/s42256-019-0048-x
58. Sageman, M. (2004). *Understanding terror networks*. University of Pennsylvania Press.
59. Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature, 461*(7260), 53–59. https://doi.org/10.1038/nature08227
60. Schrodt, P. A. (2014). Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research, 51*(2), 287–300. https://doi.org/10.1177/0022343313499597
61. Scrivens, R., Davies, G., & Frank, R. (2020). Measuring the evolution of radical right-wing posting behaviors online. *Deviant Behavior, 41*(2), 216–232. https://doi.org/10.1080/01639625.2018.1556994
62. Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton University Press.
63. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3319–3328). PMLR.
64. United Nations General Assembly. (2006). *The United Nations Global Counter-Terrorism Strategy* (A/RES/60/288).
65. United Nations General Assembly. (2015). *Plan of Action to Prevent Violent Extremism: Report of the Secretary-General* (A/70/674).
66. United Nations Security Council. (2017). *Resolution 2396* (S/RES/2396).
67. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.
68. Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research, 47*(4), 363–375. https://doi.org/10.1177/0022343309356491
69. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., & Weinberger, K. (2019). Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6861–6871). PMLR.
70. Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems 32* (pp. 9240–9251).
71. Zivot, E., & Andrews, D. W. K. (1992). Further evidence on the Great Crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics, 10*(3), 251–270. https://doi.org/10.1080/07350015.1992.10509904