Forecasting political instability using machine learning methods

Abstract

This study leverages machine learning to forecast political instability, focusing on high-risk states, using the Political Instability Task Force (PITF) dataset. Employing logistic regression with an expanding window validation strategy, the research evaluates predictive performance across one-, two-, and five-year horizons, addressing the challenge of rare events through the Area Under the Precision-Recall Curve (AUPRC) metric. Key predictors include regime type, infant mortality, years of stability, and regional conflict spillover. The model achieves robust performance, with mean AUPRC values of 0.166, 0.148, and 0.190 for the respective horizons, demonstrating stable discriminative ability (ROC-AUC 0.76). Feature importance analysis, using logistic regression coefficients and SHAP values, reveals that regional conflicts and factional democracies significantly drive instability risks, as evidenced in case studies of Pakistan (2003) and Venezuela (2016). By balancing predictive accuracy with interpretability, this study offers actionable insights for policymakers, extending prior work through multi-horizon forecasting and enhanced feature selection. Despite challenges with delayed data and a limited set of predictors, this study highlights the value of clear, interpretable machine learning models for preventing conflicts and opens the door to developing more advanced, multimodal early warning systems in the future.

Contents

1 Introduction

Political instability, marked by a state's vulnerability to disruptions such as coups, revolutions, civil wars, or social unrest, poses significant risks to domestic and international stability. Instability in one country can trigger humanitarian emergencies, economic devastation, and international security crises, highlighting the urgency of predicting such disruptions, especially as global political risk has reached a five-year high [18].

In International Relations (IR) research, the deductive paradigm has dominated, with scholars building theoretical frameworks and testing them against historical cases. For instance, realist theories, such as those advanced by Waltz, propose that state behavior is driven by the anarchic structure of the international system, a hypothesis tested against events like the Cold War rivalry between the United States and the Soviet Union. Such approaches prioritise theory-building over empirical investigation, aiming to fit complex realities into preconceived models [26].

This study adopts a positivist worldview, viewing instability as a measurable phenomenon amenable to systematic analysis through objective data, such as economic or political indicators. In contrast, interpretive approaches, such as social constructivism and critical theory, argue that instability is shaped by historical and cultural contexts, resisting standardized predictions. While these perspectives offer valuable qualitative insights, their predictive power is limited compared to data-driven methods [12].

This research aims to identify states at risk of instability, particularly those defying international norms, often termed "rogue states," which are prone to destabilizing actions like supporting terrorism or human rights abuses. Advances in Machine Learning (ML) offer new opportunities to improve forecasting, enabling preventive measures, such as diplomacy or economic aid, as emphasized in the UN-World Bank report on early warning systems. Unlike traditional methods relying on expert judgment and subjective qualitative analysis, ML leverages computational power to identify patterns in diverse datasets, providing objective, scalable, and timely predictions to support conflict prevention [25, 10, 4, 24].

ML models, grounded in mathematical and statistical principles, serve as rigorous frameworks with measurable precision. Researchers can start with informed assumptions—e.g., that economic or demographic indicators predict instability—and refine them based on empirical findings. For example, Exploratory Data Analysis (EDA) examines datasets for relationships, such as correlations between economic decline and instability risk, guiding the selection of predictive models like linear regression for interpretable results.

A key challenge in ML is balancing predictive accuracy with model interpretability. Complex algorithms may obscure causal pathways, producing "black-box" predictions. This research addresses this trade-off through careful model selection and simpler statistical methods, ensuring forecasts are accurate and comprehensible for policymakers [10, 21].

Methodologically, this study uses a country-year dataset from the Political Instability Task Force (PITF), compiled by Baillie et al. The analysis begins with EDA to uncover descriptive statistics and variable relationships, followed by modeling with validation technique to assess predictive accuracy at one-, two-, and five-year horizons. Key predictors will be analyzed to understand their role in instability risks [1].

The study is centered on the following research questions:

- What is the predictive performance of logistic regression in forecasting political instability using PITF data?
- Which features contribute most significantly to the model's predictions of political instability, and what is the nature of their impact?
- How does the predictive accuracy of the model change when forecasting instability over different time horizons (1 year, 2 years, and 5 years)?
- To what extent can the model offer interpretable explanations for its predictions, and how can these insights be applied to policy-making?

2 Literature Review

This chapter discusses the feasibility of predicting political instability using machine learning, its methodological foundations, possible benefits, and intrinsic challenges. It draws a distinction between explanatory and predictive political science models, assesses the prospects for predicting instability, and addresses the constraints that render it challenging to do so.

2.1 Explanation vs Prediction

The application of predictive approaches has become more popular in political science, especially peace and conflict research, to forecast instability and guide prevention measures. Hegre et al. [10] conceptualise forecasting as making predictions about unrealised results based on realised data. This leads to the key question: how is prediction different from explanation, and why is this difference important? Shmueli [23]'s discussion in ”To Explain or to Predict?” explains this distinction. Explanatory modelling tests hypotheses in order to establish causality—typical of IR in connecting factors such as economic inequality to unrest—whereas predictive modelling prioritizes empirical accuracy in predicting upcoming events, e.g., political instability, without requiring deep causal understanding.

This study follows a predictive approach, applying the PITF dataset to an inductive, data-driven methodology. It forecasts instability at one-, two-, and five-year time frames, focusing on accuracy over theoretical adherence. Shmueli [23] warns against confusing explanatory power and predictive accuracy—a pitfall this study avoids by prioritizing models that improve forecasting over causal complexity, a trade-off justified by its policy uses. Gleditsch [6] supports this pragmatic stance, noting that policy relevance often demands clear, actionable predictions over theoretical purity, a perspective that aligns with this study's focus on empirical utility. Whereas explanatory models require theoretically consistent variables (e.g., state legitimacy), predictive models require large, generalisable datasets like the PITF's, using cross-validation to avoid overfitting. Variable selection reflects this focus, fuelled by EDA and feature selection rather than pre-established theoretical constructs. Although Shmueli [23] comments that machine learning methods such as random forests have the potential to hide causal inference from interpretable regression models, methods like SHAP (SHapley Additive exPlanations) [16] address this by measuring feature contributions, achieving a balance between accuracy and interpretability. Ward et al. [27] endorse this approach, arguing that explanatory models' focus on statistical significance is often not followed by predictive power. They mention variables like GDP per capita as strong predictors, which concurs with this study's empirical focus. Murphy et al. [19] also appreciate interpretability in conflict forecasting, a goal this research pursues by ensuring informative outputs of the PITF data.

2.2 Prospects of prediction

Machine learning provides significant potential for forecasting political instability, enhancing theoretical understanding and practical utility. Murphy et al. [19] show how algorithmic complexity, combined with rich data and computational power, improves conflict prediction. Besides theory development, it also guides policy by improving resource allocation—e.g., peacekeeping missions to highest-risk sites—reducing costs and increasing impact. Gleditsch [6] underscores this policy linkage, arguing that predictions, even if not perfectly accurate, can inform resource allocation by highlighting likely scenarios, a benefit this study leverages through its multi-year forecasts. With high-resolution data, one can monitor in real time and respond quickly to emerging threats, surpassing static models' capabilities. This method also fosters interdisciplinary collaboration between political science and data science along with economics. The inclusion of various variables—economic data, market data—improves predictive accuracy and clarifies conflict determinants, strengthening theory-building. Accurate forecasts provide actionable risk assessments, enhancing prevention and intervention decision-making [19].

The potential of machine learning in forecasting political unrest is further evidenced by practical applications like the Violence Early-Warning System (ViEWS) developed by Hegre et al. [9]. This project successfully generates transparent, publicly available forecasts of political violence across Africa, demonstrating the feasibility and utility of such predictive efforts. Alongside other advancements, ViEWS highlights ML's potential to transform instability research and management through leveraging data abundance and computational power.

2.3 Limitations of prediction

Despite its promise, the prediction of political instability encounters notable challenges. Hegre et al. [10] and Murphy et al. [19] identify data quality issues in datasets like the PITF that include missing entries (e.g., few livelihood surveys in Africa) and biases due to media coverage. Chadefaux [3] mentions government misrepresentation as another complicating factor. Rare events like coups pose modelling challenges, as their infrequency hinders the ability to identify patterns [10]. Gleditsch [6] adds that some phenomena, such as war onset in crises, may be inherently unpredictable due to random elements in human decision-making. Complex models may also not be transparent, although this research alleviates this to some extent by preferring simpler, interpretable models where possible. Chadefaux [3] argues that path dependence and strategic flexibility of wars induce inherent uncertainty, a reality this study acknowledges by making predictions probabilistic. Ethical concerns also complicate prediction: Hegre et al. [10] caution against self-fulfilling prophecies where predictions destabilise a region, whereas Chadefaux [3] observes that strategic responses by actors can render forecasts unsustainable. Murphy et al. [19] highlight the pragmatic necessity of interpretability, which is here attended to with feature importance analysis (e.g., SHAP [16]) to provide qualitative context alongside quantitative results. These limitations underscore the need for a balanced strategy that respects forecasting's boundaries while striving to enhance its usefulness.

2.4 Previous models

Efforts to predict conflict have long utilised quantitative methods, initially focusing on interstate dynamics. Schrodt [22] applied a bootstrapped ID3 algorithm to interstate conflict data (1945–1994), achieving 95–100% accuracy on known cases but only 50–60% on new cases, representing an early application of machine learning. Brandt et al. [2] advanced this with Bayesian time series models (BVAR and MS-BVAR) for real-time interstate conflict forecasting (1996–2008), focusing on dynamic escalation patterns, though both approaches remained centered on interstate rather than intrastate phenomena.

Goldstone et al. [8] analysed global instability onsets (1955–2003), including both civil wars and democratic reversals, using a case-control approach with conditional logistic regression comparing 117 instability cases to 351 matched controls. Their parsimonious model, which achieved over 80% accuracy in forecasting instability two years in advance, relied on only four predictors: a nonlinear five-category regime type derived from Polity components, infant mortality, the occurrence of armed conflict in neighbouring states, and state-led discrimination. Notably, partial democracies with factionalism emerged as the strongest predictor, showing odds ratios exceeding 30. The model's robustness was further verified through out-of-sample validation for the period 1995–2004, confirming its predictive accuracy and practical utility.

Goldsmith et al. [7] developed a two-stage probit model to forecast the onset of genocide and politicide using a global dataset covering 1974–2003. In the first stage, they estimate political instability—encompassing events such as civil wars and regime shifts—which then informs the second-stage model that predicts genocide risk. Emphasising thorough testing, the model was validated out-of-sample for 1988–2003, where it achieved an outstanding ROC-AUC of 0.8878. Moreover, it correctly classified 81.8% of genocide onset cases and 82.7% of stable cases, showing strong predictive accuracy. Key factors driving the model's performance include dynamic, time-sensitive variables like political assassinations and election periods, which significantly contributed to its forecasting power.

Kennedy [13] thoroughly tested his forecasting model for state failure using an out-of-sample approach, dividing the data into a training set (1955–1995) and a test set (1996–2012). His findings show that a standard unconditional logistic regression applied to the full dataset yielded robust performance—achieving out-of-sample AUC values around 0.812—without significant improvements from more complex sampling techniques like case-control under-sampling. Moreover, his evaluation method incorporated ROC curve analysis and a cost-sensitive decision boundary, showing that his model's precision remains effective under realistic policy constraints (specifically, the model's threshold is optimal only if the cost of a false positive is less than 7.7% of the cost of non-intervention).

Hegre et al. [9] evaluated their ViEWS political violence forecasting system using a strict out-of-sample testing strategy, partitioning historical data (up to 2014) for training/calibration and using a distinct subsequent period (2015–2017) as the unseen test set. Assessing performance primarily with ROC-AUC and AUPRC, their ensemble models demonstrated high accuracy on the test set. Specifically for state-based conflict during 2015–2017, the system achieved an out-of-sample ROC-AUC of 0.956 and AUPRC of 0.869 at the country level, and an ROC-AUC of 0.948 and AUPRC of 0.277 at the subnational (PRIO-GRID) level, substantially outperforming baseline models.

Baillie et al. [1] developed explainable models for political instability using logistic regression with only three predictors (polity code, infant mortality, years of stability). They employed an out-of-sample testing strategy, training models on pre-2000 data and validating them on a distinct, unseen test period (reference years 2000–2015). Performance was primarily evaluated using AUPRC due to event rarity, achieving out-of-sample AUPRC values of 0.078 (1yr), 0.084 (2yr) on the test set, verifying the models' robustness despite their simplicity. Their emphasis on explainability revealed polity code (especially factional democracy) as often dominant, though infant mortality drove predictions in cases like Cote d'Ivoire (2000), supported by counterfactual analysis. I would like to thank Baillie et al. [1] for publicly sharing their cleaned PITF dataset on the Open Science Framework (osf.io/3gr72), which enabled me to replicate and build upon their findings in my coursework. Table 1 provides a concise summary of key results from some of these previous models.

Table 1: Summary of Previous Models

Authors	Test years	Metric
Goldstone et al. (2010)	1995–2004	Accuracy > 80%
Goldsmith et al. (2013)	1988–2003	ROC-AUC = 0.8878
Kennedy (2015)	1996–2012	ROC-AUC ≈ 0.812
Hegre et al. (2019)	2015–2017	ROC-AUC=0.956, AUPRC=0.869
Baillie et al. (2021)	2000–2015	AUPRC=0.078 (1yr), 0.084 (2yr)

3 Methodology

3.1 Data Description and Exploratory Data Analysis

This study utilises the PITF dataset, compiled by Baillie et al. [1], which provides a cleaned and preprocessed country-year record of political instability events from 1949 to 2016, with standardised features. The dataset includes variables such as polity codes (measuring regime type), infant mortality rates, years of stability, life expectancy, GDP per capita, etc. As the data has been pre-processed by Baillie et al. [1], it is free of major inconsistencies, though limitations like potential reporting biases or missing data in conflict zones are still noted and considered during analysis. For a comprehensive overview of the dataset structure, variable definitions, and coding procedures, readers are referred to the "Dataset and Coding Guidelines" provided by the PITF [17], which is included in the references to this paper.

To gain a deeper understanding of the dataset and its suitability for forecasting political instability, an EDA was conducted. One of the initial and most striking observations is the distribution of the target variable—political instability events. Figure 1 illustrates the frequency of political instability occurrences versus stable periods across the country-year records from 1949 to 2016. The graph reveals a pronounced class imbalance, with instances of political instability being significantly rarer than periods of stability. Specifically, political instability events, which include revolutionary wars, ethnic wars, adverse regime changes, and genocides/politicides, constitute only a small fraction of the dataset compared to the overwhelming majority of stable country-years. This disparity is not unexpected, as political instability is inherently a rare phenomenon relative to the prevailing norm of political stability across nations and time.

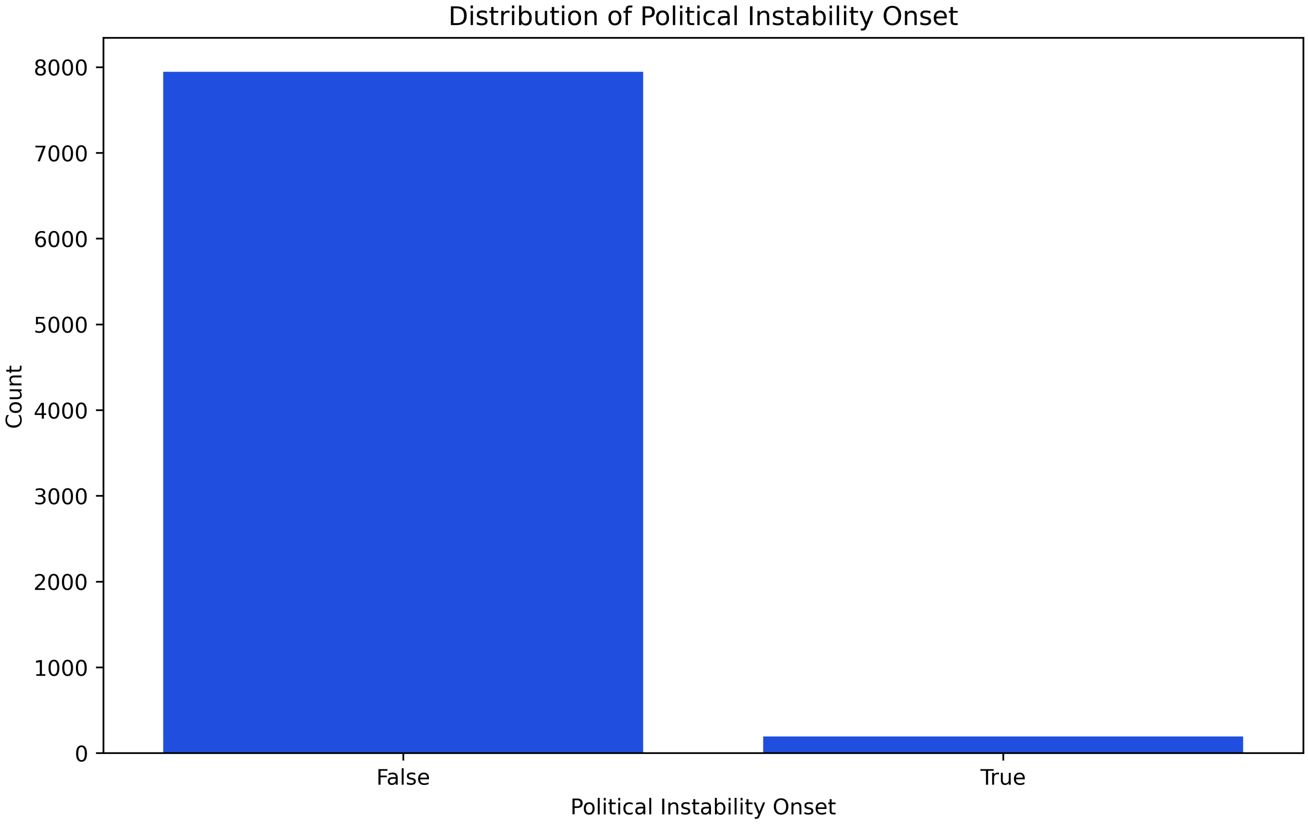

Figure 1: Frequency Distribution of Political Instability vs. Stable Country-Years (1949-2016)

This pronounced class imbalance poses a critical challenge for predictive modelling, as standard metrics like accuracy can be misleading—models may achieve high accuracy simply by predicting the majority class (stability) without effectively identifying the rare instances of instability. To address this issue and better reflect model performance in the context of imbalanced data, two evaluation metrics are employed in this study: the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). AUPRC is selected as the primary metric due to its focus on the trade-off between precision (the proportion of correctly predicted instability events among all predicted instabilities) and recall (the proportion of actual instability events correctly identified), making it particularly suitable for assessing the model's ability to detect the minority class (instability) while minimising false positives. In contrast, ROC-AUC provides a broader measure of the classifier's overall ability to discriminate between stable and unstable country-years, capturing its performance across both classes. Combining these metrics ensures a comprehensive evaluation, with AUPRC emphasising the detection of rare but critical political instability events and ROC-AUC offering insight into the model's general discriminative power.

Having established the rarity of political instability events and the associated challenges for predictive modelling, the next step in the EDA was to examine the composition of these events by type. Figure 2 presents a bar chart illustrating the count of instability episodes for each type over the given period. The results highlight a clear disparity in the frequency of different instability types. Adverse regime changes are the most common, with the highest count of episodes, totalling approximately 100. This prevalence can be attributed to the broad definition of adverse regime changes in the PITF dataset, which encompasses major and abrupt shifts from open electoral systems to authoritarian regimes, revolutionary changes in political elites, contested dissolutions of federated states, secessions, and complete or near-total collapses of central authority [17]. Such a wide scope captures a variety of governance failures, from military coups to state disintegrations, making this category the most frequent.

Ethnic wars are the second most frequent type, with around 80 episodes, driven by conflicts between governments and ethnic minorities seeking autonomy or status changes, reflecting tensions in diverse societies. Revolutionary wars follow with about 50 episodes, involving violent attempts to overthrow governments, a narrower criterion explaining their lower frequency. Genocides and politicides are the least common, with approximately 30 episodes, due to their extreme nature as systematic extermination events, often linked to ethnic or revolutionary conflicts.

This distribution highlights the heterogeneity of instability events, with adverse regime changes dominating due to their broad definition, while genocides/politicides are rare due to their extremity. These findings suggest that predictive models may need to account for type-specific characteristics through feature engineering, particularly given the prevalence of adverse regime changes and ethnic wars in hotspots like Africa and Eastern Asia & Oceania.

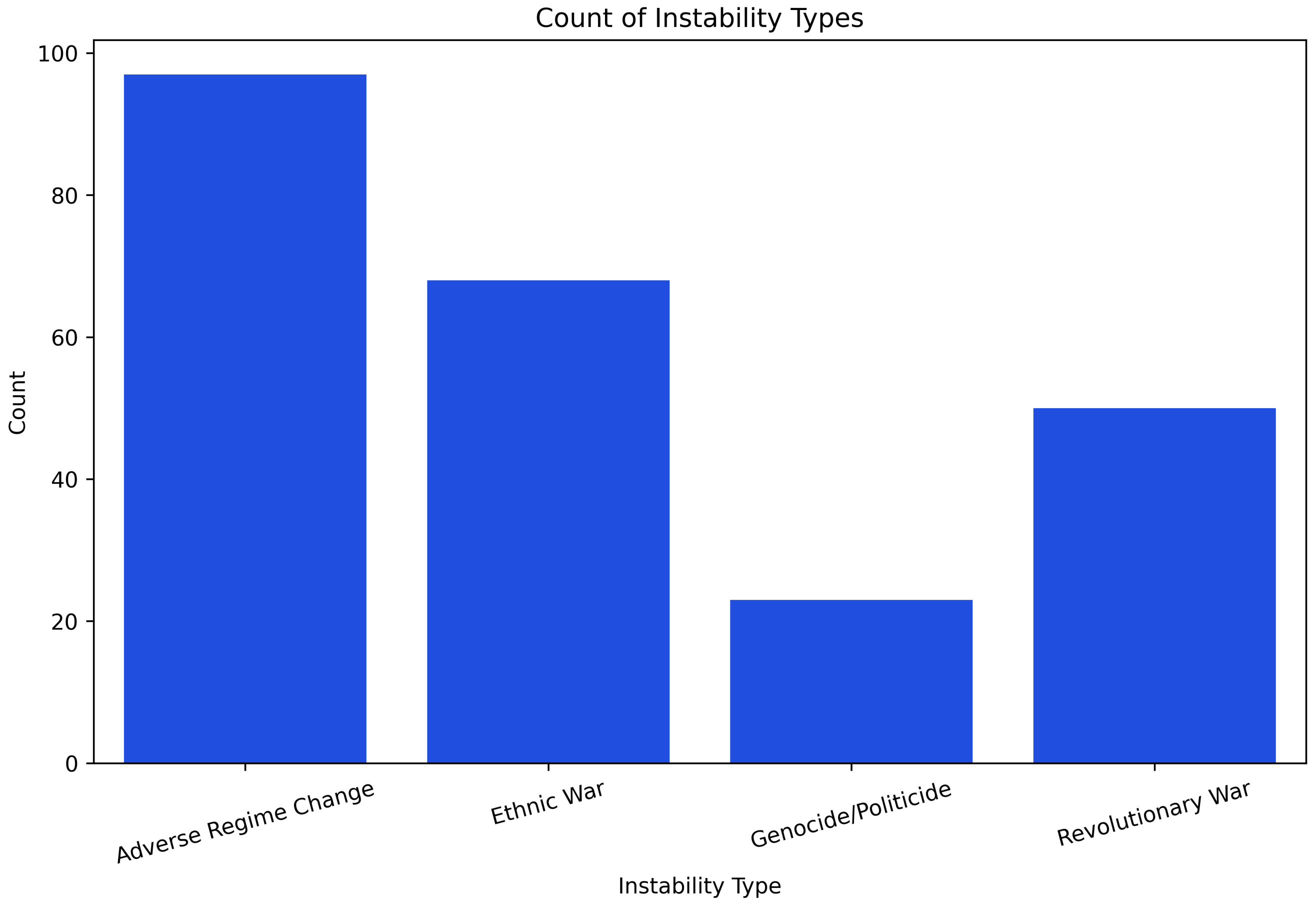

Figure 2: Count of Political Instability Episodes by Type (1949-2016)

Following the analysis of class distribution, the geographic variation in political instability events was examined to identify potential regional patterns. Figure 3 presents the proportion of country-years marked by political instability across major world regions, as defined in the PITF dataset, over the period from 1949 to 2016. The results highlight significant regional disparities in the prevalence of instability. Africa and Eastern Asia & Oceania account for the majority of instability events, collectively representing the highest shares of recorded cases. This is followed by the Middle East & Central Asia, which also exhibits a notable concentration of instability, though to a lesser extent. The Americas rank next, with a moderate incidence of political instability, while Europe stands out as the region with the lowest proportion of instability events.

These findings align with historical trends and contextual factors, such as post-colonial conflicts, ethnic diversity, and resource struggles in Africa, as well as revolutionary movements and territorial disputes in Eastern Asia & Oceania. The relatively low incidence in Europe may reflect greater political stability following World War II, bolstered by economic and political integration. This regional variation suggests that geographic context could be a valuable feature in predictive modelling, potentially improving the model's ability to differentiate between stable and unstable country-years when combined with socio-economic and political indicators.

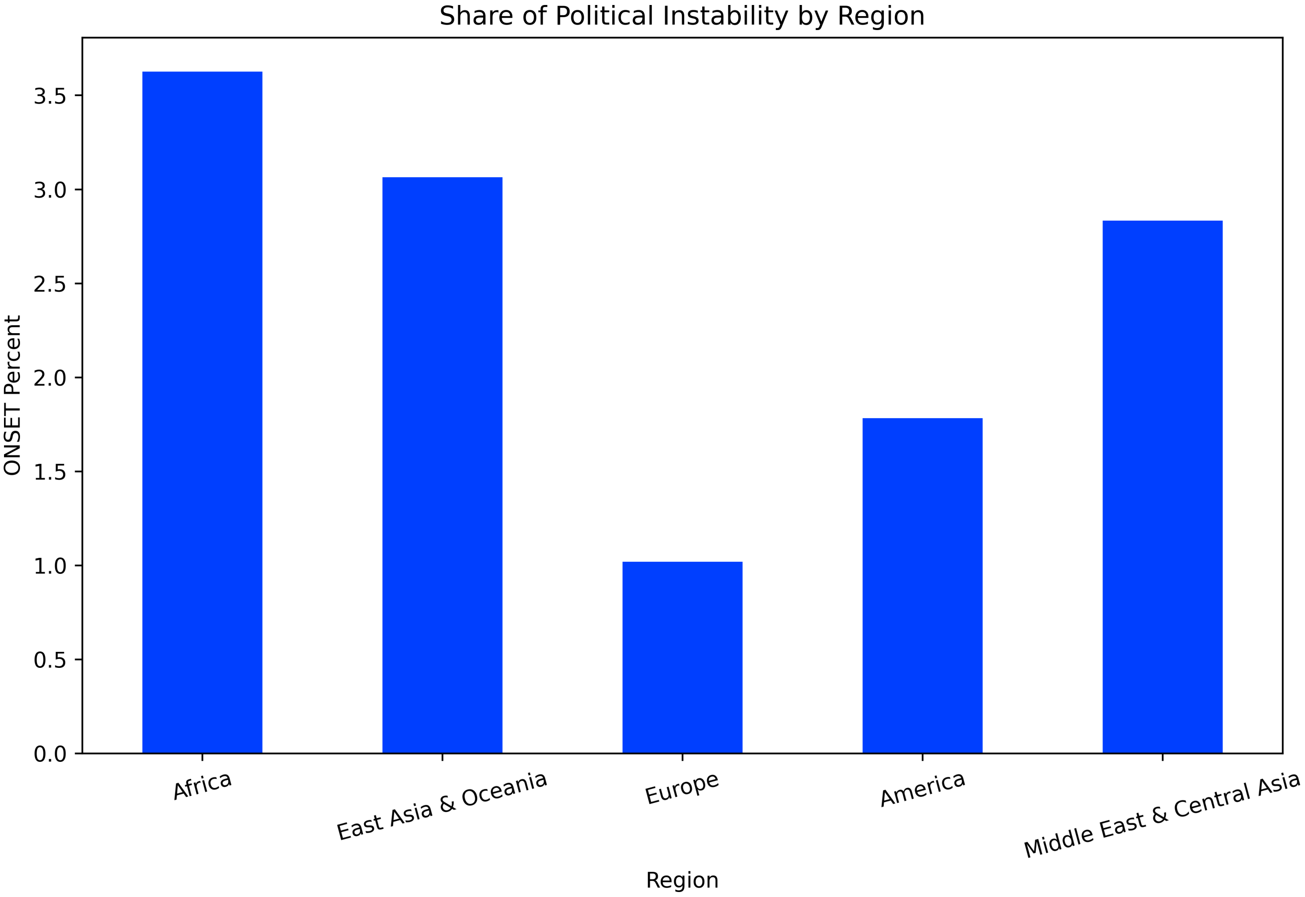

Figure 3: Proportion of Unstable Country-Years by Region (1949-2016)

To complement the regional analysis, the temporal distribution of political instability events was examined. Figure 4 illustrates the annual proportion of country-years marked by the onset of political instability, highlighting how the prevalence of such events has evolved over time. The trend reveals significant fluctuations, with distinct peaks and periods of decline. In the immediate post-World War II years (1949–1960), the proportion of instability events remains relatively low, with a slight increase toward the late 1950s. This is followed by a series of pronounced spikes in the 1960s and 1970s, where the proportion of instability events frequently reaches 5%–6%, likely reflecting the wave of decolonisation in Africa and Asia, as well as Cold War-related conflicts in various regions. A notable peak occurs around 1990, with the proportion of instability events reaching approximately 7%, potentially driven by the dissolution of the Soviet Union and ensuing conflicts in Eastern Europe and Central Asia. After 2000, the trend shows a general decline, with the proportion of instability events dropping to near-zero levels by 2015, though minor fluctuations persist.

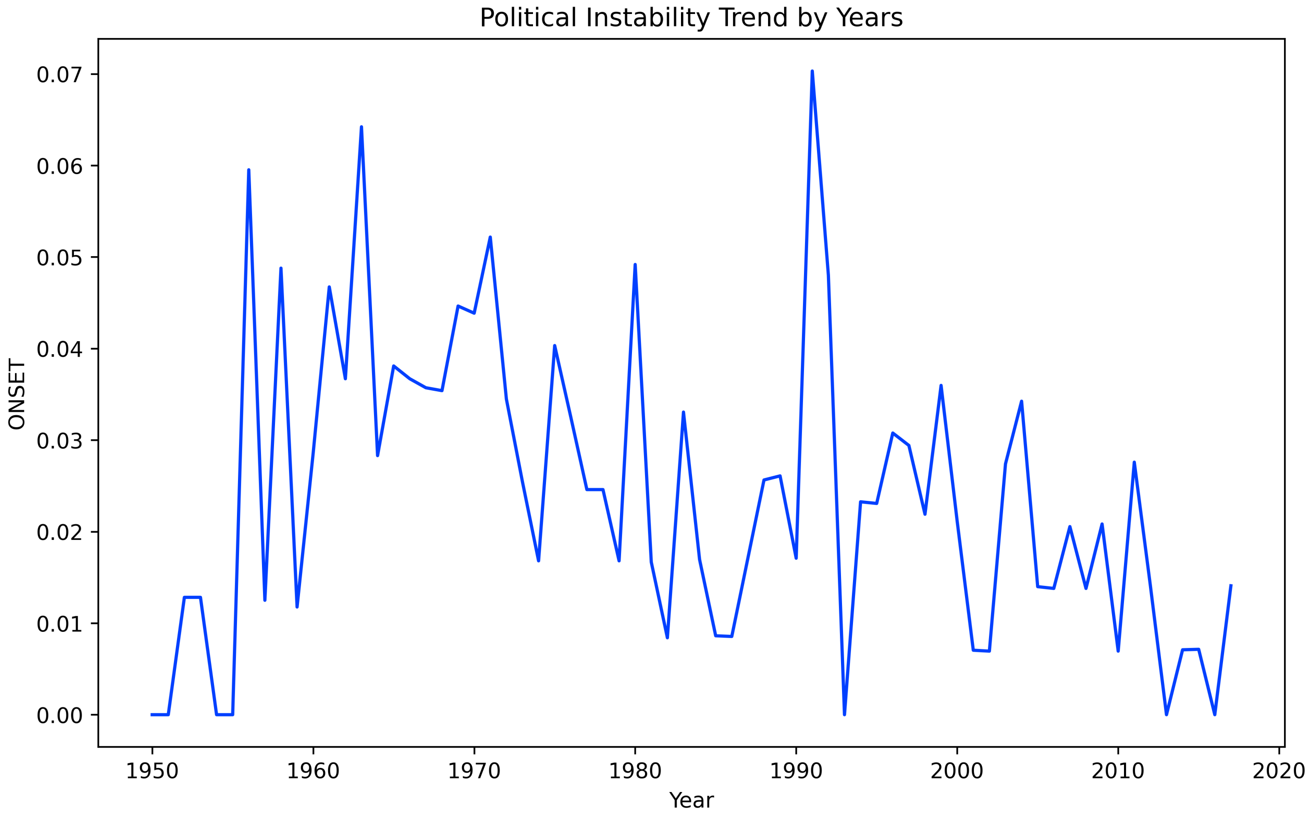

Figure 4: Annual Proportion of Unstable Country-Years (1949-2016)

The EDA provided critical insights into the PITF dataset, laying a robust foundation for forecasting political instability. The analysis revealed a significant class imbalance, with political instability events being rare compared to stable country-years, necessitating the use of the AUPRC as the primary evaluation metric to ensure effective detection of these rare events. Regionally, Africa and Eastern Asia & Oceania emerged as hotspots for instability, while Europe exhibited the lowest incidence, suggesting that geographic context is a key factor in predictive modelling. Temporally, the data showed distinct peaks in instability during the 1960s, 1970s, and around 1990, driven by historical events such as decolonisation, Cold War conflicts, and the dissolution of the Soviet Union, followed by a decline post-2000. The distribution of instability types highlighted the predominance of adverse regime changes and ethnic wars, attributed to their broader definitions and the prevalence of governance failures and communal tensions, particularly in high-risk regions. These findings underscore the importance of incorporating regional, temporal, and type-specific features into the predictive model to capture the complex dynamics of political instability. For a more detailed analysis, including the code used to generate the visualisations and additional exploratory steps, readers are referred to the GitHub repository linked in the appendix.

3.2 Testing Strategy: Expanding Window Validation

Following the EDA, which established the characteristics and challenges of the PITF dataset, a predictive model was constructed and validated to forecast political instability. To evaluate the model's performance in a manner that mirrors its intended real-world application, an expanding window validation strategy was employed, covering the period from 2000 to 2017. This approach was selected for its ability to simulate the practical scenario of forecasting future instability based solely on historical data available at the time of prediction, ensuring that the model's performance reflects realistic conditions faced by policymakers.

The expanding window validation technique was implemented as follows. The PITF dataset, as structured by Baillie et al. [1], contains three key temporal columns: `DATAYEAR` (the year of the country data), `REFYEAR` (the reference year from which the forecast is made), and `EVENTYEAR` (the year for which political instability is predicted, occurring or not). In this structure, `DATAYEAR` is always one year prior to `REFYEAR`, and `EVENTYEAR` is one year after `REFYEAR`, reflecting a one-year-ahead forecasting setup. The dataset was split into a series of training and testing periods, with the training window incrementally expanding each year from 2000 to 2016. For each test year t (ranging from 2000 to 2016), the model was trained on all country-year observations where `DATAYEAR` was less than t (i.e., historical data available up to t - 1), and its predictions were tested on records where `REFYEAR` equaled t, targeting instability events in the corresponding `EVENTYEAR` (t + 1). This process replicates a real-world forecasting scenario: for example, in 2025, a policymaker would use data up to 2024 (`DATAYEAR`) to predict instability in 2026 (`EVENTYEAR`), with the forecast made from 2025 (`REFYEAR`), as 2025 data would not yet be fully available. In this validation, the first split trained the model on data up to `DATAYEAR` 1999 to predict instability in `EVENTYEAR` 2001 (tested on `REFYEAR` 2000), the second split used data up to `DATAYEAR` 2000 for `EVENTYEAR` 2002 (tested on `REFYEAR` 2001), and so forth, concluding with training on data up to `DATAYEAR` 2015 for `EVENTYEAR` 2017 (tested on `REFYEAR` 2016). The detailed implementation of this process is provided in the code attached in the appendix.

This approach generated 17 distinct train-test splits, with the training sample size growing progressively from 5,701 country-years in the first split (training until 1999 for testing on 2000) to 7,996 country-years by the final split (training until 2015 for testing on 2016). The test sample size remained relatively consistent, ranging from 140 to 146 country-years per split, reflecting the number of countries observed in each year, with minor variations due to data availability. The 2000–2016 period was chosen for validation not only because it aligns with the post-2000 decline in instability events observed in the EDA (Figure 4), but also to test the model's performance on more recent examples. This focus on later years ensures an assessment of the model's relevance to contemporary contexts rather than relying solely on older historical data, providing a rigorous evaluation of its ability to detect rare events in a relatively stable global environment and its robustness to temporal shifts in instability patterns.

The expanding window validation strategy offers distinct advantages for this study. Notably, it allows the model to incorporate an increasing volume of historical data with each iteration, enhancing its capacity to capture evolving long-term trends, such as the post-2000 decline in instability or regional variations identified in the EDA. Additionally, this method ensures that predictions are made in a forward-looking manner, aligning with the operational needs of policymakers who require timely and actionable forecasts based on the most recent data available.

To assess model performance across these splits, two metrics were calculated for each test year: AUPRC and ROC-AUC. Greater emphasis was placed on AUPRC as the primary evaluation metric, given the pronounced class imbalance in the dataset—instability events constitute only approximately 2% of country-years, as noted in the EDA. AUPRC was prioritised because it effectively measures the model's ability to detect these rare positive instances (instability events) by balancing precision and recall, making it more suitable than accuracy or ROC-AUC alone for identifying the minority class in this imbalanced context. ROC-AUC was included as a secondary metric to provide a broader assessment of the model's discriminative power, capturing its overall ability to distinguish between stable and unstable country-years across both classes. However, certain test years contained only stable country-years with no instability events. These years were excluded from the evaluation because AUPRC and ROC-AUC cannot be meaningfully calculated without at least one instability event present to compare against stable cases. Consequently, only years with both stable and unstable outcomes were included. Final performance was determined by averaging the AUPRC and ROC-AUC scores across all eligible test years, providing a robust estimate of the model's generalisation ability over the 2000–2016 period. This validation framework thus lays a solid foundation for subsequent steps, such as feature selection and model optimisation, ensuring that the forecasting approach remains both empirically sound and practically relevant.

3.3 Model Selection and Configuration: Logistic Regression

Following the EDA and the establishment of the expanding window validation strategy, the subsequent phase involved selecting and configuring an appropriate machine learning model for forecasting political instability. The primary objectives guiding model selection were twofold—achieving high predictive accuracy, particularly in identifying the rare instances of political instability, and ensuring model interpretability to understand the factors driving the predictions.

Several candidate models were considered and evaluated, including Logistic Regression (LR), Decision Trees, Random Forests, and Gradient Boosting algorithms. Experimental results indicated that while ensemble methods, specifically Gradient Boosting, yielded slightly higher performance based on the primary evaluation metric (AUPRC), Logistic Regression provided a compelling balance between predictive capability and interpretability. Given that interpretability is a key requirement for this study—enabling insights into why certain country-years are flagged as high-risk—the minor predictive edge of Gradient Boosting was deemed less critical than the transparency offered by LR. This decision aligns with the methodological choice of Baillie et al. [1], who also utilised Logistic Regression in their analysis of the same PITF dataset, further supporting its suitability for this task.

Logistic Regression is a statistical model well-suited for binary classification problems, such as predicting the presence or absence of political instability. It models the probability of the outcome variable (instability event = 1, stable = 0) as a function of the predictor variables (features like polity score, infant mortality, GDP per capita, etc.). The model outputs a probability score, and its coefficients can be interpreted to understand the direction and magnitude of influence each feature has on the likelihood of instability, fulfilling the interpretability requirement.

A critical consideration, highlighted during the EDA, is the pronounced class imbalance within the PITF dataset, where instability events represent a small fraction (2-3%) of the total country-years. Standard model training can lead to poor performance on the minority class. To mitigate this, the Logistic Regression model was specifically configured to address the imbalance by setting the class_weight hyperparameter to ’balanced’. This crucial adjustment modifies the learning process by assigning higher weights to the minority class (instability events) and lower weights to the majority class (stable periods), effectively forcing the model to pay more attention to correctly classifying the rare but critical instability instances. This internal weighting mechanism complements the choice of AUPRC as the primary evaluation metric, ensuring that both the model training and its evaluation are focused on effectively identifying political instability despite its rarity. Consequently, this configured Logistic Regression model was carried forward for implementation and assessment using the expanding window validation strategy.

3.4 Feature Selection

Feature selection for this study involved choosing the most effective predictors from the comprehensive set of variables available in the pre-processed PITF dataset compiled by Baillie et al. [1].

The selection process was initially guided by the findings of Baillie et al. [1], who highlighted regime type, standardised log infant mortality, and standardised years of stability as particularly decisive predictors in their analysis. Taking these core concepts as a starting point, this study evaluated their predictive contribution alongside other variables readily available in the Baillie et al. [1] dataset. The goal was to determine if incorporating other pre-existing features could enhance predictive performance within this study’s specific validation framework (expanding window, AUPRC metric).

To enhance the model, this study integrated the strengths of Baillie et al. [1] with insights from [8]. Specifically, the evaluation revealed that adding the `nc4` feature—a binary indicator for whether 4 or more neighboring states have armed conflict (¡span class=”math-inline”¿ geq 4¡/span¿ neighboring states in conflict)—which was utilised by Goldstone et al. [8] — significantly improved model performance (as measured by AUPRC) compared to using only the three core features recommended by Baillie et al. [1].

Consequently, the final set of features selected from the Baillie et al. [1] dataset for the Logistic Regression model comprises:

1. `pol_code`: The categorical variable representing regime type.

2. s_logim: The standardised logarithm of the infant mortality rate (pre-processed).

3. s_StabilityYears: The standardised count of consecutive years of stability (pre-processed).

4. nc4: The binary indicator for whether 4 or more neighboring states have armed conflict.

All four selected variables were directly available in the dataset prepared by Baillie et al. [1]. The only data transformation performed in this study was the application of one-hot encoding to the pol_code variable to prepare it for the Logistic Regression model. This final four-feature combination demonstrated the strongest empirical performance within this study's validation setup.

4 Results

This section presents the predictive performance results for the Logistic Regression models developed to forecast political instability over different time horizons: 1-year, 2-year, and 5-year periods. All models utilised the same feature set (pol_code, s_logim, s_StabilityYears, nc4) and algorithm configuration (Logistic Regression with class_weight='balanced'). They were evaluated using the expanding window validation strategy detailed in the Methodology, covering reference years from 2000–2016 for out-of-sample testing. Performance was primarily assessed using AUPRC and secondarily using ROC-AUC.

The models differ in their target variable definition: The 1-year model predicts instability onset in the next year (EVENTYEAR = REFYEAR + 1). The 2-year model predicts onset within the next two years (i.e., an event occurs in REFYEAR + 1 OR REFYEAR + 2). The 5-year model predicts onset within the next five years (i.e., an event occurs in any year from REFYEAR + 1 through REFYEAR + 5).

The overall model performance for each horizon was determined by averaging the yearly scores across all valid evaluation splits within the 2000–2016 reference period. As noted in the Methodology, reference years where the test set contained only stable outcomes relative to the specific forecast horizon were excluded. For the 1-year model, this applied to reference years 2012 and 2015 (resulting in 15 valid years included in the average). However, for the 2-year and 5-year models, instability events occurred within their respective forecast windows for all reference years; therefore, their performance metrics were averaged across all 17 yearly splits (2000–2016).

Table 2 summarises the average performance metrics (Mean ± Standard Deviation) for each model horizon:

Table 2: Summary of Model Performance Metrics

Forecast Horizon	Valid Years Averaged	Mean ROC-AUC ± SD	Mean AUPRC ± SD
1-Year	15	0.760 ± 0.141	0.166 ± 0.250
2-Year	17	0.771 ± 0.090	0.148 ± 0.156
5-Year	17	0.767 ± 0.062	0.190 ± 0.092

The ROC-AUC scores indicate consistently good discriminative ability across all horizons, peaking slightly for the 2-year model and showing progressively lower year-to-year variability (smallest SD) as the horizon increases. The AUPRC scores, crucial for evaluating performance on the rare instability events, show the highest average performance for the 5-year horizon (0.190). Intriguingly, the 2-year model shows a slightly lower average AUPRC than the 1-year model, despite being averaged over more years. Similar to ROC-AUC, the variability (standard deviation) of AUPRC significantly decreases as the forecast horizon lengthens, indicating more stable year-to-year precision-recall performance for longer-term predictions.

Direct comparison of metrics, particularly AUPRC, across different time horizons should be made with caution due to the change in the underlying prediction task. Extending the forecast window increases the baseline probability of an instability event occurring within that window for any given country-year flagged as 'at-risk'. This higher prevalence of "positives" within the longer window means the model has more opportunities over time to correctly register an event, which can make achieving higher and more stable AUPRC scores potentially easier compared to the more demanding task of pinpointing instability onset within a single year.

The higher AUPRC for the 5-year horizon reflects the relative ease of predicting instability over a longer period. As the forecast window extends, the baseline probability of an instability event increases (e.g., from 2-3% for a 1-year horizon to a higher rate over 5 years), allowing the model to capture more events within the prediction period. Consequently, predicting instability within 5 years is less challenging than forecasting it for the next year, as the model benefits from a greater likelihood of events occurring and more stable year-to-year performance, as evidenced by the lower variability in AUPRC.

Figures 5, 6, and 7 display the year-by-year performance trends (AUPRC and ROC-AUC) for the 1-year, 2-year, and 5-year models, respectively. Note that the plot for the 1-year model excludes data points for 2012 and 2015. These visualisations illustrate the performance dynamics and variability specific to each forecast horizon.

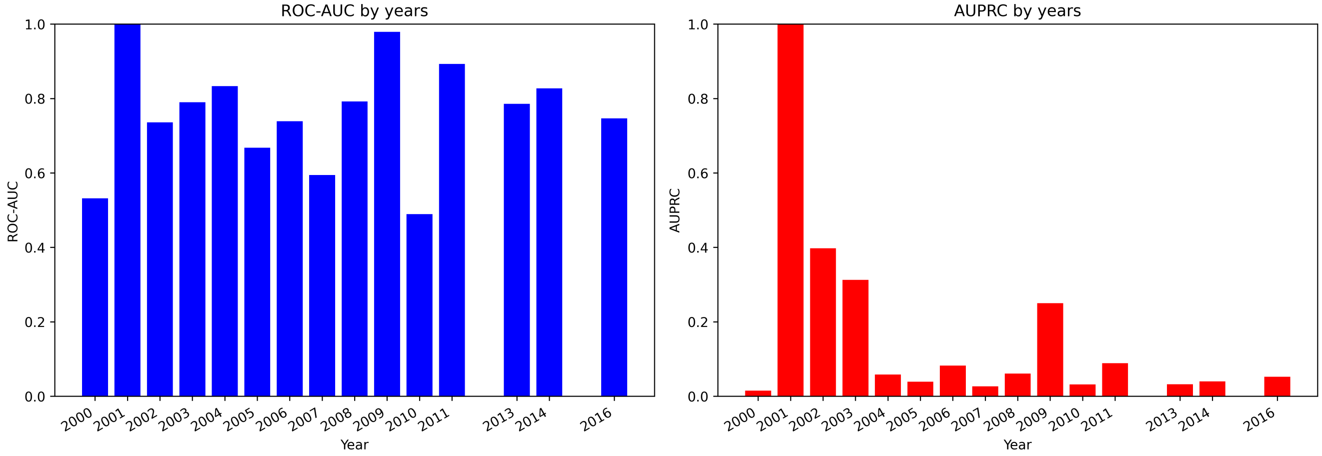

Figure 5: Year-by-Year Performance (AUPRC and ROC-AUC) for the 1-Year Forecast Horizon (2000-2016)

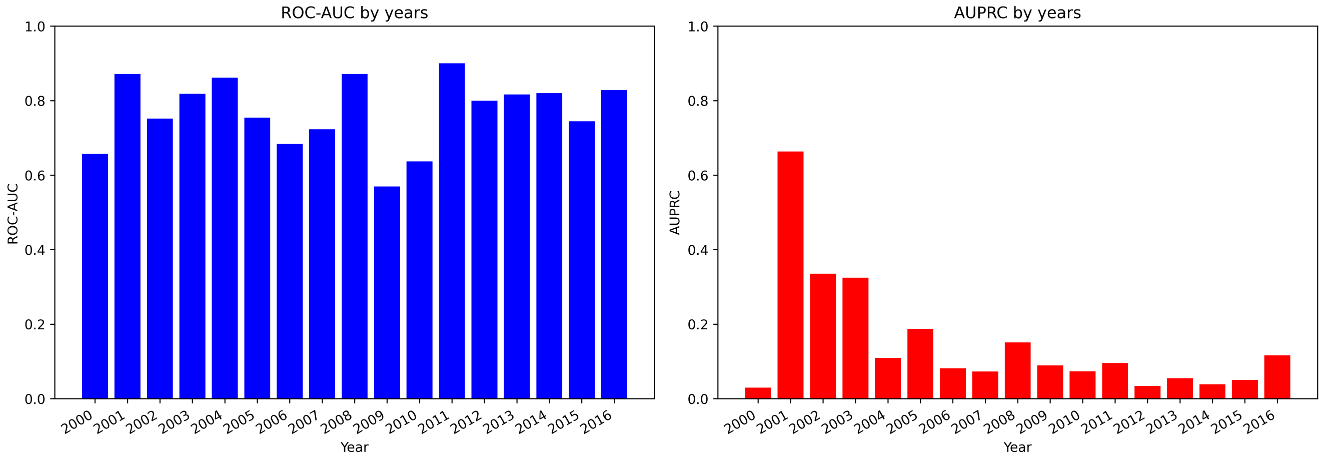

Figure 6: Year-by-Year Performance (AUPRC and ROC-AUC) for the 2-Year Forecast Horizon (2000-2016)

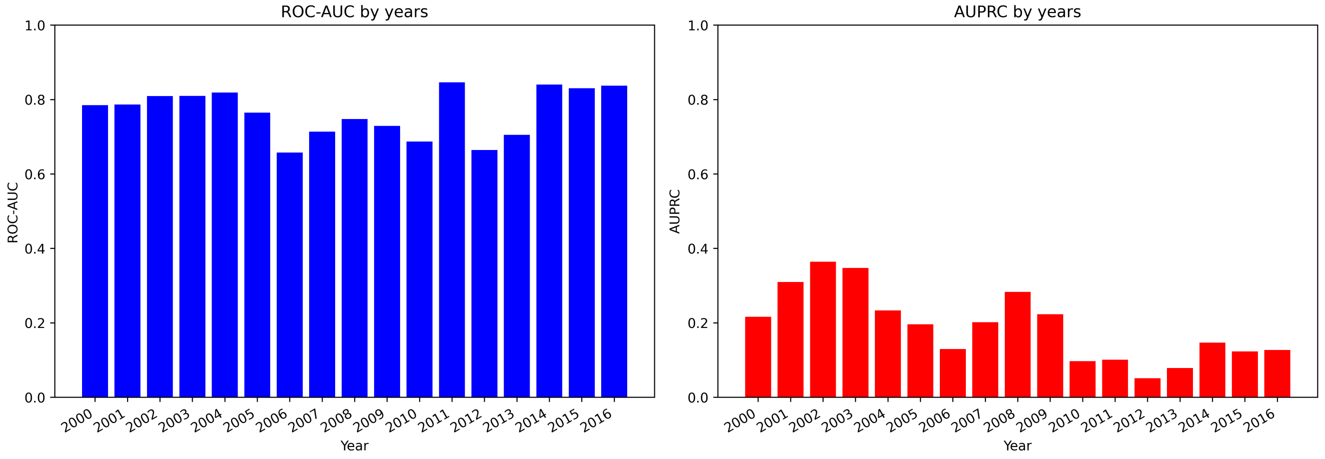

Figure 7: Year-by-Year Performance (AUPRC and ROC-AUC) for the 5-Year Forecast Horizon (2000-2016)

5 Feature Importance Analysis

Following the evaluation of the Logistic Regression models' predictive performance, this section investigates the contributions of the selected features (`pol_code`, `s_logim`, `s_StabilityYears`, `nc4`) to the forecasting of political instability. The analysis employs a dual approach to clarify the drivers of instability: a global assessment of feature importance through the model's coefficients captures their overall predictive roles across the Political Instability Task Force (PITF) dataset, while a local analysis using SHAP (SHapley Additive exPlanations) examines their impact on specific instability cases. This methodology aligns with the study's commitment to balancing predictive accuracy with interpretability, as emphasised in the Methodology section, and extends the interpretable modelling approaches of [8] and [1]. By integrating broad patterns with context-specific insights, the analysis enhances the understanding of instability dynamics and supports the development of early-warning systems.

5.1 Global Feature Importance using Logistic Regression Coefficients

To understand the overall impact of each predictor variable on the likelihood of political instability across the entire dataset, a global feature importance analysis was conducted using the coefficients of the trained Logistic Regression model. In Logistic Regression, the coefficient for a feature indicates the change in the log-odds of the positive outcome (political instability) for a one-unit increase in that feature, holding all other features constant. The magnitude of the coefficient (its absolute value) reflects the strength of the relationship, while the sign indicates the direction (positive coefficients increase the odds of instability, negative coefficients decrease them). The odds ratio, calculated as the exponential of the coefficient ($\exp(\beta)$), provides a more intuitive interpretation: it represents the factor by which the odds of instability change for a one-unit increase in the feature.

The analysis of the Logistic Regression coefficients for the 1-year forecast model (trained within the expanding window framework) yielded the following insights into global feature importance, sorted by the absolute magnitude of the coefficients:

Table 3: Logistic Regression Model Coefficients and Odds Ratios

Feature	Coefficient (β)	Odds Ratio
nc4	1.369083	3.931742
pol_code_dem	-1.245447	0.287812
pol_code_fdem	1.152021	3.164581
s_logim	0.704683	2.023205
pol_code_auc	-0.219180	0.803177
pol_code_pdem	0.097890	1.102841
pol_code_pauc	-0.032041	0.968466
s_StabilityYears	-0.031690	0.968807

The feature with the largest absolute coefficient, nc4 (involvement in $\geq$ 4 external conflicts), shows the strongest overall positive association with political instability (coefficient 1.369, odds ratio 3.932). This means that, holding other factors constant, involvement in multiple external conflicts is associated with nearly a four-fold increase in the odds of political instability. This finding is highly significant and aligns with established understanding in international relations regarding the spillover and destabilizing effects of external conflicts on domestic stability [8].

The pol_code features, representing different regime types, are also prominent drivers of instability, as indicated by their coefficients. The signs and relative magnitudes of these coefficients clearly reveal which regime types are associated with higher or lower risk. The strongest associations are seen with pol_code_dem (democracy), which has a large negative coefficient (-1.245), indicating a strong association with decreased instability risk, and pol_code_fdem (factional democracy), which has a large positive coefficient (1.152), indicating a strong association with increased instability risk. This stark contrast highlights that while established democracies are generally the most stable regime type in this analysis, factional democracies are significantly more unstable, corroborating findings from previous studies that emphasize the particular vulnerability of partial or factional democratic systems [8, 1].

Other regime types show less pronounced associations. pol_code_auc (autocracy) has a negative coefficient (-0.219), suggesting a slight decrease in instability risk. pol_code_pdem (partial democracy) has a small positive coefficient (0.098), indicating a minimal association with increased risk. pol_code_pauc (partial autocracy) has a coefficient very close to zero (-0.032), indicating almost no unique association with instability risk in this model when other factors are considered.

The s_logim feature (standardised log infant mortality) is the fourth most important feature globally, with a positive coefficient (0.705) and an odds ratio of 2.023. This suggests that higher infant mortality rates are strongly associated with increased odds of political instability, with each standard deviation increase in infant mortality approximately doubling the odds of instability. As a proxy for public health, socio-economic conditions, and state capacity, this variable consistently emerges in the literature as a key indicator of fragility [1, 8].

The s_StabilityYears feature (standardised years of stability) has a very small negative coefficient (-0.032) and an odds ratio close to 1 (0.969). While longer periods of stability are associated with a slight decrease in the odds of instability, its global impact as measured by the Logistic Regression coefficient is relatively minor compared to the regime type, external conflict, or infant mortality.

Figure 8 visually represents these coefficients, illustrating the magnitude and direction of each feature's association with the predicted log-odds of political instability. The features with coefficients furthest from zero, such as nc4, pol_code_dem, and pol_code_fdem, are globally the most influential predictors in this Logistic Regression model.

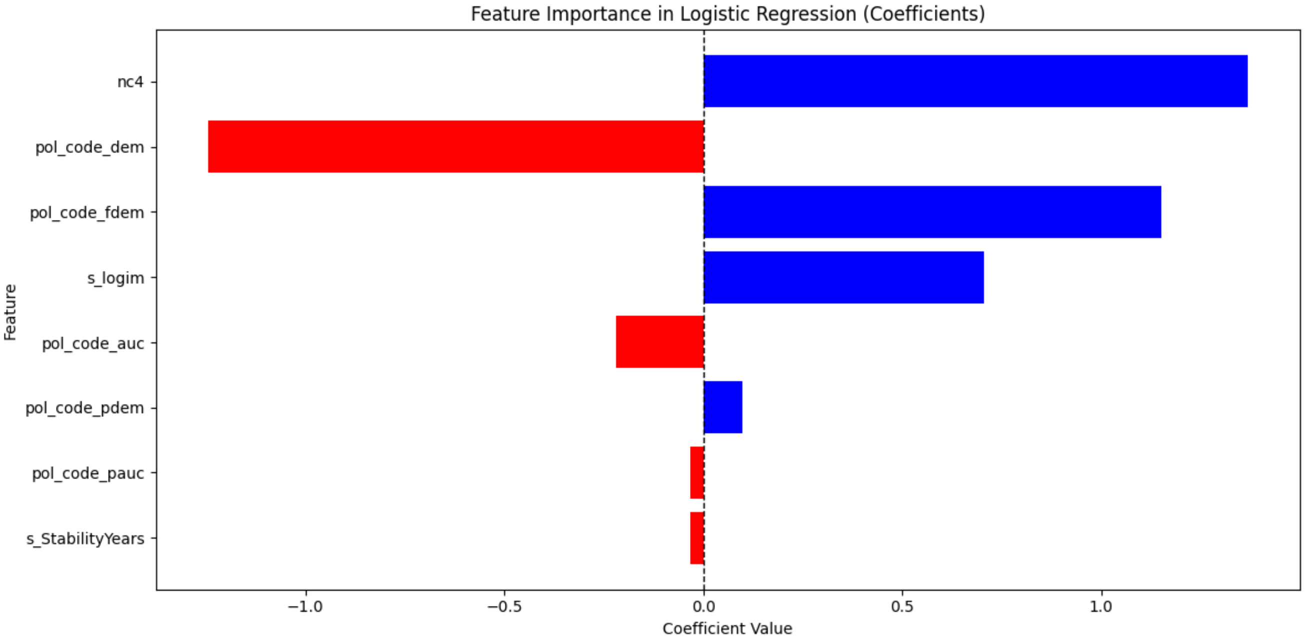

Figure 8: Visual Representation of Logistic Regression Coefficients

This global analysis using coefficients provides a high-level overview of which features the model relies on most heavily when making predictions across the entire dataset. It confirms that factors related to external conflict, specific regime types (especially the contrast between democracy and factional democracy), and socio-economic conditions (infant mortality) are the primary global drivers of political instability in this model. However, this global perspective does not reveal how these features interact or influence predictions for specific country-year cases, which necessitates the local analysis using SHAP values discussed in the following sub-section.

5.2 Local Feature Importance Analysis using SHAP

To complement the global feature importance derived from Logistic Regression coefficients, SHAP (SHapley Additive exPlanations) values are used to examine how individual features influence predictions of political instability for specific country-year cases [16]. This section outlines how SHAP values are calculated and interpreted in the logistic regression model.

The logistic regression model predicts the log-odds of political instability for a given example (e.g., a country in a specific year). The log-odds are calculated as a linear combination of features (1):

$$\text{log-odds}(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{1}$$

where β_0 is the intercept, and $(\beta_1, \ldots, \beta_p)$ are coefficients for features $(x_1, \ldots, x_p)$.

SHAP values break down this log-odds prediction into contributions from each feature, measured relative to a baseline (the average log-odds across the training set). For feature i, the SHAP value ϕ_i quantifies its contribution based on its coefficient β_i and how its specific value x_i deviates from the average $\bar{x}_i$ (2):

$$\phi_i = \beta_i(x_i - \bar{x}_i) \tag{2}$$

where β_i: the coefficient of feature i, x_i: the value of feature i, $\bar{x}_i$: the mean value of feature i.

The predicted log-odds for a specific case is then the sum of the baseline log-odds and the SHAP values for all features (3):

$$\text{log-odds}(\mathbf{x}) = \phi_0 + \phi_1 + \phi_2 + \cdots + \phi_p \tag{3}$$

where $\phi_0 = \beta_0 + \sum_{i=1}^{p} \beta_i \bar{x}_i$ is the baseline log-odds. Interpreting SHAP values is straightforward: a positive ϕ_i indicates that feature i's value for this specific case increases the predicted log-odds of political instability (pushing the prediction towards instability), while a negative ϕ_i indicates it decreases the log-odds (pushing the prediction towards stability). The magnitude of ϕ_i shows the strength of this push.

SHAP values thus clarify how features like years of stability or regime types shape predictions for specific cases. The following sections will analyze SHAP values for selected country-year examples to highlight their contributions to predicted instability risks.

5.3 Explaining the case of Pakistan, 2003

This case study analyzes Pakistan (CCODE=770) in 2003, where the model, trained on data up to 2003, predicted a high probability of political instability (0.827) for 2004, accurately capturing the onset of the Insurgency in Khyber Pakhtunkhwa. In 2004, tensions in Pakistan's Federally Administered Tribal Areas (FATA), particularly in Waziristan, escalated into armed conflict when the Pakistan Army's operations to apprehend al-Qaeda operatives, fleeing from Afghanistan, provoked resistance from local militant groups, including the Tehrik-i-Taliban Pakistan (TTP) and their allies [15]. These operations, aligned with the U.S.-led War on Terror, fueled local grievances and militant mobilization in response to cross-border instability. SHAP values elucidate the drivers behind the model's prediction, as illustrated in Figure 9.

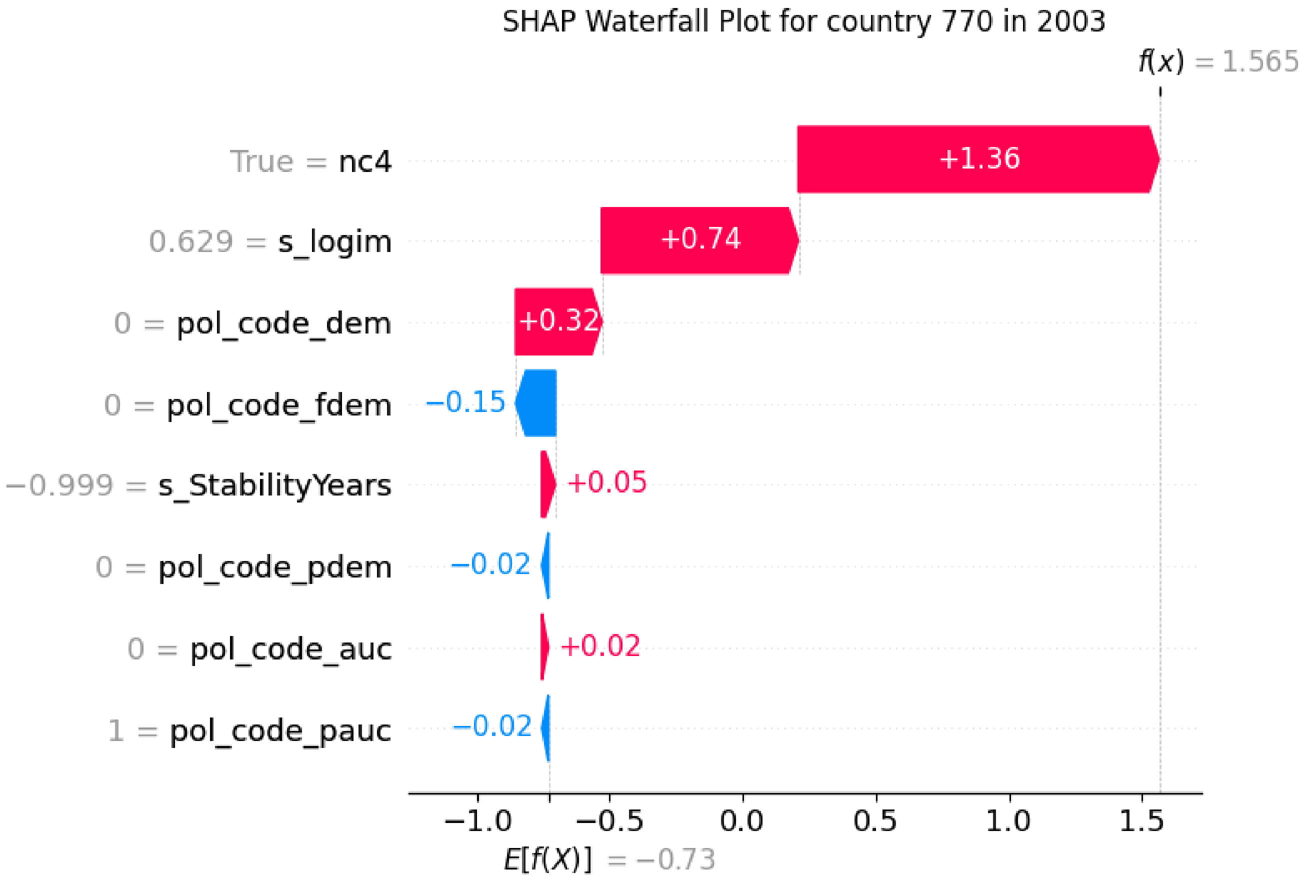

Figure 9: SHAP Waterfall Plot for Pakistan, 2003

The model's predicted log-odds for Pakistan was 1.565, starting from a baseline log-odds of −0.73. The most significant contribution came from `nc4` ($\phi_{\text{nc4}} = 1.355$), indicating that four or more neighboring states were engaged in armed conflicts (`nc4` = `True`). This reflects the spillover effects from the ongoing war in Afghanistan, which amplified militant activities and instability in Pakistan's FATA region. The second-largest contribution was from `s_logim` ($\phi_{\text{s\_logim}} = 0.739$), with a standardized log infant mortality rate of 0.629, signaling above-average socio-economic deprivation.

The absence of a democratic regime (`pol_code_dem` = 0, $\phi_{\text{pol\_code\_dem}} = 0.324$) amplified the instability risk, as the lack of democracy removed a stabilizing factor. The feature `s_StabilityYears` ($\phi_{\text{s\_StabilityYears}} = 0.051$) also contributed positively, as its value of −0.999 indicated fewer years of stability than the average. This reflects Pakistan's recent history of political volatility, slightly increasing the log-odds of instability.

Other features had smaller impacts. The partial autocracy regime (`pol_code_pauc` = 1, $\phi_{\text{pol\_code\_pauc}} = -0.024$) and absence of factional democracy (`pol_code_fdem` = 0, $\phi_{\text{pol\_code\_fdem}} = -0.150$) had negligible effects, as did `pol_code_pdem` and `pol_code_auc` ($|\phi| < 0.025$). These minor contributions suggest that regime type played a secondary role compared to regional conflict spillover and socio-economic conditions in this case.

5.4 Explaining the case of Venezuela, 2016

This case study examines Venezuela (CCODE=101) in 2016, where the model, trained on data up to 2016, predicted a probability of political instability (0.570) for 2017, correctly identifying the onset of widespread anti-government protests. In early 2017, the Supreme Court, aligned with President Nicolás Maduro, attempted to dissolve the opposition-controlled National Assembly, triggering a constitutional crisis and mass demonstrations across the country [11]. These protests, fueled by hyperinflation, shortages of food and medicine, and accusations of authoritarian governance, demanded immediate presidential elections and the release of political prisoners. Security forces responded with excessive force, resulting in dozens of deaths and thousands of arrests between April and September 2017. SHAP values reveal the drivers behind the model's prediction, as shown in Figure 10.

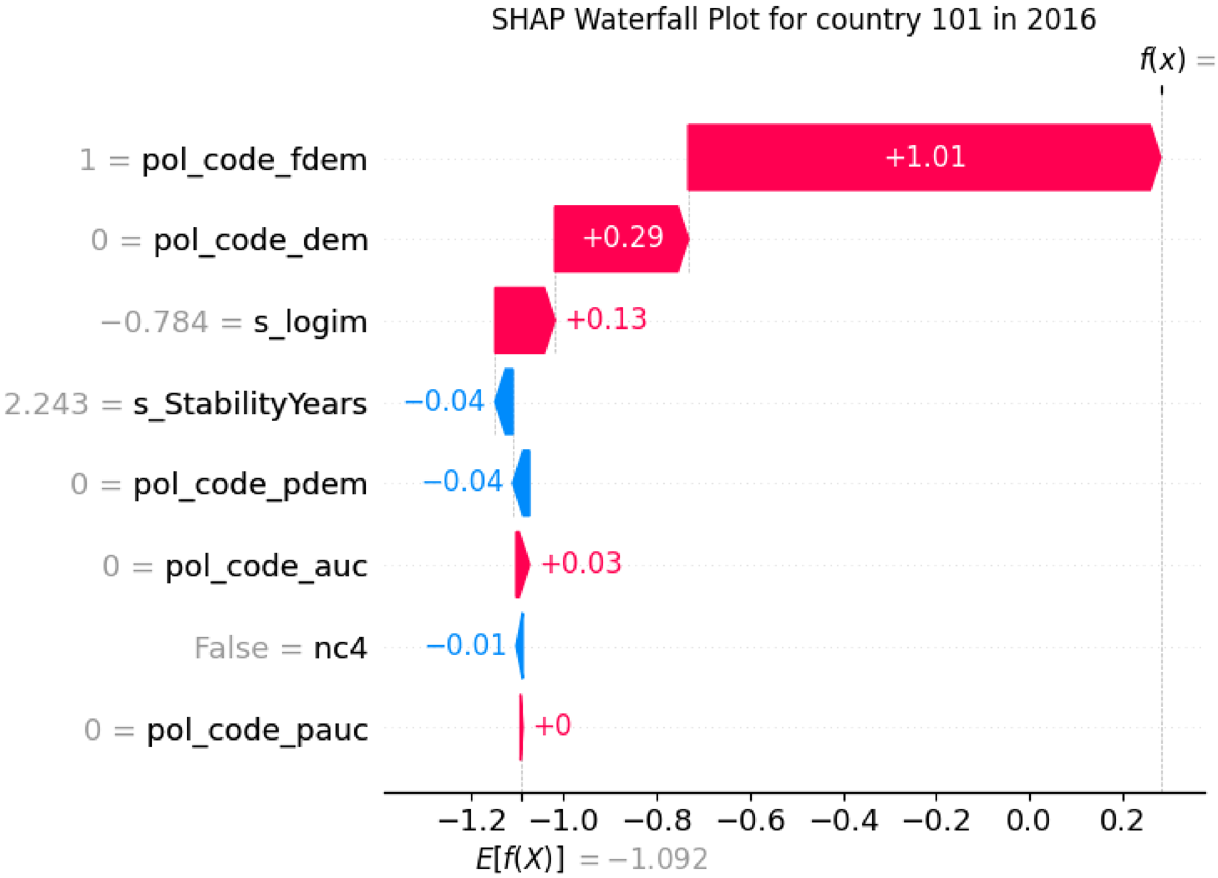

Figure 10: SHAP Waterfall Plot for Venezuela, 2016

The model's predicted log-odds for Venezuela was approximately 0.282, starting from a baseline log-odds of −0.73. The largest contribution came from `pol_code_fdem` ($\phi_{\text{pol\_code\_fdem}} = 1.014$), indicating the presence of a factional democracy (pol_code_fdem = 1). This regime type, characterized by a mix of democratic institutions and authoritarian practices, significantly increased the risk of instability, reflecting Venezuela's polarized political environment under Maduro's rule. The absence of a full democratic regime (pol_code_dem = 0, $\phi_{\text{pol\_code\_dem}} = 0.286$) further amplified the log-odds, as the lack of democracy removed a stabilizing factor.

Other features, including s_logim, s_StabilityYears, nc4, and regime type dummies (pol_code_pdem, pol_code_auc, pol_code_pauc), had smaller contributions ($|\phi| < 0.14$), suggesting that socio-economic conditions, regional conflicts, and specific regime characteristics played a secondary role compared to political polarization and governance issues in this case.

6 Discussion

6.1 Results interpretation

The results of this study build on the foundational work of Goldstone et al. [8] and Baillie et al. [1], integrating their strengths to enhance predictive performance and interpretability. By combining key predictors from both studies—polity code and infant mortality from Goldstone et al. [8], and polity code, infant mortality, and years of stability from Baillie et al. [1]—this model captures both structural and temporal drivers of political instability. The inclusion of conflict in neighboring states, inspired by Goldstone et al. [8], further enriches the model by accounting for regional spillover effects, leading to improved performance across the test years compared to using Baillie et al. [1] predictors alone. However, direct comparisons with prior studies are approached cautiously due to differences in test years, validation techniques, and evaluation metrics, which preclude definitive claims of superior accuracy.

A key contribution of this study lies in its enhanced model interpretability, drawing on Baillie et al. [1] emphasis on transparency. Global feature importance analysis revealed that polity code, particularly factional democracy, consistently emerged as the dominant predictor across high-risk cases, aligning with findings from both Goldstone et al. [8] and Baillie et al. [1]. Local case explanations, inspired by Baillie et al. [1] counterfactual approach, provided nuanced insights into specific predictions, enabling identification of whether polity code, infant mortality, or other factors drove instability in individual country-years. This dual interpretive framework enhances the model's practical utility, offering policymakers clear, actionable insights into the drivers of instability.

Unlike prior studies, this model extends the forecasting framework to include 1-, 2-, and 5-year horizons, addressing a gap in Goldstone et al. [8] two-year focus and Baillie et al. [1] one- and two-year models. The consistent performance across these horizons, particularly the novel 5-year window, underscores the model's robustness in capturing both short-term triggers and longer-term trends. Additionally, by adopting AUPRC as a performance metric, as recommended by Baillie et al. [1], the model effectively addresses the challenge of forecasting rare events (2-3% base rate in PITF data), ensuring reliable rankings of high- and low-risk country-years.

6.2 Limitations

This study developed and evaluated an interpretable model for forecasting political instability using the PITF dataset and a specific set of features. However, several limitations related to the data, the chosen methodology, and the forecasting task itself should be acknowledged.

First, the analysis relies exclusively on the pre-processed Political Instability Task Force (PITF) dataset provided by [1]. Although this dataset benefits from cleaning and standardisation, it is still subject to the underlying limitations of the original PITF data collection. This includes potential gaps in coverage or data quality, particularly for countries experiencing active conflict or those with authoritarian regimes where reliable information access is restricted [10, 19]. Furthermore, potential reporting biases (e.g., from media coverage or official sources) and inconsistencies in variable measurement across diverse country contexts over time could influence the data used for training and testing [3, 19]. While the [1] dataset addresses many issues, any residual noise or bias directly impacts the performance and generalisability of the model developed here.

Second, a practical limitation stems from the temporal lag associated with several key predictors used in the final model. Indicators such as the standardised log infant mortality (`s_logim`) or the count of stability years (`s_StabilityYears`) are based on demographic and political data that often become available with a significant delay (potentially 1–2 years). Similarly, the `nc4` indicator depends on the timely reporting and coding of external conflicts. This inherent data latency means that forecasts, although generated using the most recently available historical data within the expanding window framework, may not capture the impact of very recent changes or rapidly developing crises, limiting real-time applicability.

Third, the scope of predictors was deliberately constrained to four features (`pol_code`, `s_logim`, `s_StabilityYears`, `nc4`) selected from the [1] dataset based on prior literature and empirical performance within this study's framework. This parsimonious approach enhances interpretability but necessarily excludes potentially valuable information from other variables within the PITF dataset or from alternative data streams altogether, such as real-time news analysis or granular geospatial data. The reliance on this specific set of established structural indicators might also limit the model's ability to capture novel or rapidly changing drivers of instability, a challenge highlighted by the observed decrease in effectiveness of some forecasting systems post-2011 [14].

Finally, as established in the EDA and addressed through the choice of AUPRC and class weighting, political instability remains an inherently rare event. Predicting the specific timing and nature of such infrequent occurrences is fundamentally challenging. Although the model demonstrates good discriminative ability (indicated by ROC-AUC) and performs reasonably well in identifying rare events (indicated by AUPRC), especially over longer horizons, it is better understood as a tool for assessing relative risk levels across countries and time rather than a system guaranteeing the prediction of every instability event. The broader context of forecasting challenges, particularly in potentially shifting global political landscapes [14], reinforces the need for cautious application.

7 Conclusion and Future Research

This study demonstrates the utility of interpretable machine learning models using established structural indicators for forecasting political instability across different time horizons. However, it represents just one component within the broader, evolving landscape of conflict prediction. The limitations discussed previously, particularly concerning data latency and the scope of predictors, point towards significant avenues for future research aimed at developing more comprehensive, timely, and robust early warning systems.

A promising frontier lies in the creation of Multimodal Early Warning Systems (MEWS). Such systems would move beyond reliance solely on traditional macroeconomic, demographic, and political indicators (like those used in this study and many systems reviewed by [20]) to integrate data from diverse sources and modalities. This could include:

- **Textual Data**: Real-time analysis of news articles, social media posts (Twitter/X, Facebook, etc.), and communication platforms (e.g., Telegram channels) using Natural Language Processing (NLP) techniques like topic modeling, sentiment analysis, and the detection of keyword surges or hate speech to gauge public mood and identify escalating narratives.

- **Event Data**: Systematically incorporating geolocated event datasets (e.g., ACLED, GDELT, UCDP GED, or mission-specific data like SAGE mentioned by [19]) to track protests, attacks, arrests, and other critical incidents as they occur.

- **Geospatial and Movement Data**: Leveraging satellite imagery and potentially anonymized mobile network data to monitor population displacement, troop movements, infrastructure changes, or environmental stress factors that can precede or correlate with instability.

- **Structured Data**: Continuing to use the foundational macroeconomic, political (e.g., Polity, V-Dem), and socio-demographic indicators, while actively seeking ways to mitigate issues of data latency and improve cross-national comparability.

The central future research challenge resides in the effective data aggregation and fusion of these heterogeneous data streams. How can structured time-series data be meaningfully combined with unstructured text, geospatial information, and discrete event data within a single predictive framework? This necessitates exploration into advanced machine learning techniques, potentially including sophisticated ensemble methods, deep learning architectures designed for multimodal input (e.g., incorporating attention mechanisms or graph neural networks), and methods for dynamic variable weighting based on data timeliness and reliability. Ensuring interpretability within these complex models will remain paramount [19].

Beyond model development, significant practical implications arise concerning deployment and operationalization. A truly effective MEWS requires transitioning from retrospective analysis to real-time operation. This involves substantial engineering efforts to build robust, scalable data ingestion pipelines, real-time feature extraction, model inference, and user-friendly interfaces for policymakers and analysts. Maintaining such systems demands continuous monitoring, evaluation, and adaptation.

Furthermore, as highlighted by [20] in their review of existing CEWS, transparency and accessibility are crucial. Future MEWS development should prioritize open standards where possible, clear documentation of data sources and methodologies, and mechanisms for independent evaluation to foster trust and collaborative improvement within the research and policy communities. Addressing ethical considerations related to data privacy (especially for movement and social media data) and the potential misuse of predictions [5, 19] must be integral to system design.

In conclusion, while significant challenges remain, the development of integrated, multimodal, and interpretable early warning systems offers the potential to significantly enhance our ability to anticipate, understand, and ultimately mitigate political instability. This study, by providing a robust and interpretable baseline using structural data, serves as a foundational step upon which these more complex and potentially impactful future systems can be built.

Code and Data Availability

The code used for the analyses presented in this study is publicly available on GitHub at the following address: `https://github.com/Zakmar2004/Political-Instability-Forecasting`

The processed dataset used for this study is the Political Instability Task Force (PITF) dataset, cleaned and shared by [1]. This dataset is publicly available for download from the Open Science Framework (OSF) repository at: `https://osf.io/3gr72/`

Declaration on the Use of Generative Models

In accordance with NRU HSE regulations, the assistance of generative models was used in the preparation of this written work.

Parts of the Text Prepared Using Generative Models: Generative models were used as an auxiliary tool in working on the text of **all sections** of this work. The models were not applied for the full or partial generation of substantive content; their primary role was in editorial support and improving the form of presentation.

Purposes and Method of Application of Generative Models: The use of generative models falls under the category of "author text with the use of generated materials". The main purposes and methods of application included:

- Assistance in structuring the text and improving the logic of presenting ideas (brainstorming).
- Correction of grammar, spelling, and punctuation of the text.
- Improvement of writing style and formulation of author's ideas for greater clarity and precision.

Generative Models Used: The following generative models were used during the preparation of this work:

- **Grok**: Available at `https://grok.x.ai/` (used for assistance with structure, brainstorming, improving style, grammar, and idea formulation).
- **Gemini**: Available at `https://gemini.google.com/` (used for LaTeX code formatting).

Assessment of the Effectiveness of Applying Generative Models: The application of generative models is assessed as **quite effective** for achieving the stated purposes. The tools proved useful for initiating ideas, enhancing the logical flow and structure of the text, and polishing the language and style of presentation.

References

[1] Baillie, E., Howe, P. D. L., Perfors, A., Miller, T., Kashima, Y., and Beger, A. (2021). Explainable models for forecasting the emergence of political instability. *PLoS ONE*, 16(7):e0254350.

[2] Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2009). Real-time, time-series forecasting of political conflict. University of Texas, Dallas; University of Minnesota; University of Kansas.

[3] Chadefaux, T. (2017). Conflict forecasting and its limits. *Journal of Peace Research*, 54(2):175–190.

[4] Colaresi, M. and Mahmood, Z. (2017). Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193–214.

[5] Druet, D. (2021). Enhancing the use of digital technology for integrated situational awareness and peacekeeping-intelligence. DPO Peacekeeping Technology Strategy.

[6] Gleditsch, K. S. (2022). One without the other? prediction and policy in international studies. *International Studies Quarterly*, 66(3):sqac036.

[7] Goldsmith, B. E., Butcher, C. R., Semenovich, D., and Sowmya, A. (2013). Forecasting the onset of genocide and politicide: Annual out-of-sample forecasts on a global dataset, 1988–2003. *Journal of Peace Research*, 50(4):437–452.

[8] Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J., and Woodward, M. (2010). A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208.

[9] 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., von Uexkull, N., and Vestby, J. (2019). Views: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174.

[10] Hegre, H., Metternich, N. W., Nygård, H. M., and Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research*, 54(2):113–124.

[11] Human Rights Watch (2017). Crackdown on dissent: Brutality, torture, and political persecution in venezuela. Report.

[12] Junjie, M. and Yingxin, M. (2022). The discussions of positivism and interpretivism. *Global Academic Journal of Humanities and Social Sciences*, 4(1):10–14.

[13] Kennedy, R. (2015). Making useful conflict predictions: Methods for addressing skewed classes and implementing cost-sensitive learning in the study of state failure. *Journal of Peace Research*, 52(5):649–664.

[14] Korotayev, A., Medvedev, I., and Zinkina, J. (2022). Global systems for sociopolitical instability forecasting and their efficiency: A comparative analysis. *Comparative Sociology*, 21(1):64–104.

[15] Lieven, A. (2017). Counter-insurgency in pakistan: The role of legitimacy. *Small Wars & Insurgencies*, 28(1):166–190.

[16] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017).*

[17] Marshall, M. G., Gurr, T. R., and Harff, B. (2019). *State Failure Problem Set: Internal Wars and Failures of Governance, 1955–2018 (Codebook, Rev. Sept. 2, 2019).* Societal-Systems Research Inc.

[18] Middleton, J. (2023). The trendline - global political risk at highest level in five years. Verisk Maplecroft.

[19] Murphy, M., Sharpe, E., and Huang, K. (2024). The promise of machine learning in violent conflict forecasting. *Data & Policy*, 6:e35.

[20] Rød, E. G., Gåsste, T., and Hegre, H. (2024). A review and comparison of conflict early warning systems. *International Journal of Forecasting*, 40(1):96–112.

[21] 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.

[22] Schrodt, P. A. (1990). Predicting interstate conflict outcomes using a bootstrapped id3 algorithm. *Political Analysis*, 2(1):31–56.

[23] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

[24] Stukal, D., Belenkov, V., and Philippov, I. (2021). Data science methods in political science research: analyzing protest activity in social media. *Political Science*, 1(1):46–75.

[25] UN-World Bank (2018). *Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict*. World Bank, Washington, DC.

[26] Waltz, K. N. (1979). *Theory of International Politics*. McGraw Hill, New York.

[27] Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.