Machine learning for humanitarian forecasting: A Survey

Assessing the trustworthiness and real-world feasibility of machine learning models for conflict forecasting

Alexia Iustina Gavrilă[1]

Supervisor(s): Dr. Cynthia Liem[1]**, Marijn Roelvink**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Alexia Iustina Gavrilă
Final project course: CSE3000 Research Project
Thesis committee: Dr. Cynthia Liem, Marijn Roelvink, Dr. Jing Sun

Abstract

As humanitarian needs increase while donor budgets decrease, anticipatory strategies are essential for effective crisis response. In this context, machine learning (ML) has emerged as a promising tool for crisis forecasting, offering the potential to support timely interventions and humanitarian decision-making. However, despite rapid developments in ML-based prediction models, questions remain about their practical utility and trustworthiness in real-world humanitarian settings. This study presents a systematic scoping review of 32 academic and gray literature sources to assess the reliability and feasibility of ML systems for conflict forecasting. By analyzing these systems across dimensions such as forecasting scope, data sources, modeling approaches, validation practices, and ethical considerations, the study finds that while some models demonstrate strong predictive performance and methodological rigor, many lack transparent validation, robust error analysis, and operational applicability. The review concludes that while ML systems hold substantial potential for enhancing conflict anticipation, their current real-world readiness is uneven and context-dependent.

1 Introduction

While the global incidence of conflicts has declined since the Cold War, the international landscape remains precarious. Hegre et al. [1] predicted a continued decrease in the proportion of countries experiencing internal armed conflict, from about 15% in 2009 to 7% by 2050. Despite this optimistic forecast, recent years have witnessed significant escalations in violence. In 2024 alone, over 120 armed conflicts were active worldwide, involving more than 60 state actors and at least 120 non-state armed groups, according to the International Committee of the Red Cross [2].

The re-emergence of interstate warfare is exemplified by the Russian invasion of Ukraine in 2022, marking the largest and deadliest war in Europe since World War II and representing a rare case of a global power pursuing territorial conquest and regime change [3]. The scale and brutality of the conflict have reshaped global security paradigms and contributed to the fact that 2022 was the deadliest year in terms of conflict-related deaths in nearly three decades since the 1994 Rwandan genocide [4]. Simultaneously, the escalation of violence in the Israel-Palestine conflict has further illustrated how longstanding geopolitical tensions can erupt into severe humanitarian crises with little warning [5].

Humanitarian organizations such as the Red Cross[1], World Bank[2], and UNICEF[3] are facing an increasing gap between the escalating need for assistance and the financial resources to meet these needs [6]. This gap has been aggravated by the increasing frequency and intensity of crises caused by conflicts, climate change, and their compounding effects [6; 7; 8]. The number of people in need of humanitarian assistance has more than doubled over the past five years [9]. However, financial and institutional resources have not kept pace. Budget cuts by major donors, including USAID[4] and the UK government[5], have further reduced the resources available to respond effectively to emergencies [10].

In light of these challenges, many humanitarian actors have started to adopt anticipatory or early-action strategies, aiming to intervene before crises escalate to minimize both human suffering and costs [11]. These strategies are not limited to natural disasters or climate-related emergencies, as they target violent conflict as well. In such contexts, accurate and timely predictions of when and where violence will occur are essential. As highlighted in a joint report by the World Bank and the United Nations[6] in 2017, early warning and early action are increasingly essential pillars in conflict prevention and humanitarian preparedness [12].

In recent years, humanitarian actors and researchers have turned to machine learning (ML) to improve predictions and responses to armed conflicts. As Cederman and Weidmann [13] stated, if ML and big data analytics “can help” us through everyday decisions, then these tools “should also be able to” predict and potentially prevent deadly conflicts. As these conflicts continue to threaten human security and global stability, ML techniques offer a promising approach to enhance rapid information collection, data analysis, and decision support. Facilitated by advancements in remote sensing, crowd-sourcing, open data availability, and enhanced computational capacities, ML could be used to support preparedness, response, and recovery in crisis situations [14].

However, translating the theoretical promise of ML into practical applications within humanitarian contexts involves significant challenges. Accurate, reliable, and trustworthy information remains essential for the decision-makers who rely on model outputs, as inaccuracies can severely impact humanitarian decisions and outcomes [15]. To better understand the opportunities and limitations of ML systems in this domain, this study conducts a systematic literature review of existing ML applications for conflict forecasting.

1.1 Recent Work

Recent research has significantly advanced our understanding of ML applications in armed conflict forecasting. Obukhov and Brovelli [16] wrote a literature review focused primarily on the conditioning factors and predictors in ML models. Their study revealed substantial variability in the predictors employed across different models, highlighting the significant role played by socioeconomic conditions and political or governance factors. However, their analysis predominantly emphasized these conditioning factors, leaving out critical aspects such as validation methods and real-world applicability of such systems in humanitarian contexts. This omission is significant because understanding how models are validated

[1] https://redcross.eu/

[2] https://www.worldbank.org/ext/en/home

[3] https://www.unicef.org/

[4] https://oig.usaid.gov/

[5] https://www.gov.uk/

[6] https://www.un.org/en/about-us

and whether they can be operationalized in practice is central to assessing their utility for real-world forecasting and decision-making.

Additionally, Rød et al. [17] conducted an extensive comparison of conflict early-warning systems, emphasizing transparency, public accessibility of data and methods, and the geographic specificity of forecasting models. Their findings showed notable inconsistencies in accuracy and practical applicability depending on regional contexts and specific conflict types. Although their review offered insights into the diversity of forecasting approaches, it still did not explicitly address the trustworthiness and real-world feasibility of ML systems within operational humanitarian settings. This gap is important because trust in model outputs is essential for humanitarian actors who must base high-stakes decisions on these forecasts.

1.2 Research Question

Recognizing these critical gaps, this research aims to conduct a systematic scoping review of ML applications in conflict forecasting, focusing on their reliability and practical feasibility in humanitarian contexts. Specifically, this study aims to answer the following research question:

> *"How reliable and feasible are machine learning systems for conflict forecasting in real-world humanitarian contexts?"*

To systematically explore this question, it will be broken down into two sub-questions:

RQ1: *"How trustworthy and accurate are existing ML models for conflict forecasting, based on their reported performance metrics and validation practices?"*

RQ2: *"Under what contextual conditions are these models practically deployable for real-world humanitarian decision-making?"*

The first sub-question examines the technical soundness of current ML models. It aims to assess how well current ML systems predict conflict events, how rigorously they are validated, and whether their reported results can be trusted for practical use. The second investigates the practical feasibility. Its goal is to assess whether the models can realistically be used in humanitarian operational contexts, considering challenges such as limited data quality, regional variability, and resource constraints.

This paper is structured as follows: Section 2 describes the methodology and search strategy used in the systematic scoping review. Section 3 presents the results of the literature analysis. Section 4 discusses key insights regarding the models' trustworthiness, feasibility, and gaps in current research. Section 5 reflects on responsible research practices. Section 6 addresses the study's limitations and outlines directions for future work. Finally, Section 7 concludes with a summary of findings and their implications for humanitarian conflict forecasting.

2 Methodology

This research employs a systematic scoping review methodology to gain insights into the current state of ML applications for conflict forecasting within humanitarian contexts, guided by the general SALSA strategy [18]. The SALSA framework comprises four main components: Search, Appraisal, Synthesis, and Analysis, providing a structured approach to systematically review the relevant literature.

2.1 Search

The first step of the SALSA framework involves the systematic search and collection of relevant literature. The searches were carried out using three academic databases: IEEE Xplore[7], Scopus[8], and Web of Science[9]. All databases were accessed and searched in early May 2025.

The search query was formulated around three key themes, each capturing an aspect of the research topic. The first part of the query includes words associated with armed conflict, such as *"war"* and *"battle"*, to ensure the inclusion of studies situated within the domain of warfare. The second part includes keywords linked to ML and artificial intelligence. The third part focuses on early warning systems, which are essential to anticipatory humanitarian responses. These three parts formed the basis of a structured search strategy for identifying relevant literature in this scoping review. Although this review is concerned with humanitarian applications, terms such as "humanitarian" were excluded from the search query after preliminary testing revealed that their inclusion significantly reduced the number of relevant results. Therefore, such terms were omitted to avoid overlooking relevant research in conflict forecasting and early warning more broadly. The exact search query used in the database queries is presented below (Query 1).

Query 1: Search for documents containing both ”armed conflict” and ”machine learning”

```
("conflict" OR "battle" OR "war" OR "armed
 conflict")
 AND
("machine learning" OR "artificial intelligence" OR
 "AI" OR "deep learning" OR "big data")
AND
("early warning" OR "early detection" OR "warning
  system" OR "alert system" OR "risk alert")
```

In total, the search yielded 494 potentially relevant publications across the three selected databases. Specifically, IEEE Xplore returned 100 publications, Scopus returned 222, and Web of Science returned 172. Additionally, 10 more publications were identified through backward and forward citation tracking, referred to as the snowballing method [19].

2.2 Appraisal

Following the literature search, the next phase of the SALSA framework is the appraisal stage, which focuses on systematically evaluating the relevance and quality of the identified publications. The goal was to ensure that only the studies that met the study's criteria were included in the final analysis.

[7] https://ieeexplore.ieee.org/Xplore/home.jsp

[8] https://www.elsevier.com/products/scopus/search

[9] https://mjl.clarivate.com/search-results

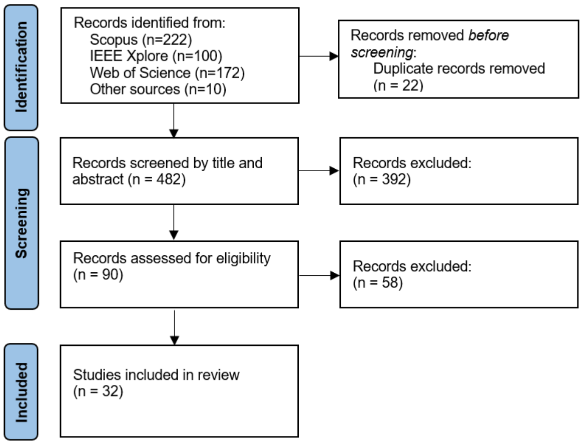

Figure 1: PRISMA Flow Diagram

The appraisal process was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, which provide a standardized protocol for systematic literature reviews [20]. Figure 1 outlines the steps followed in selecting and excluding publications using the PRISMA flow diagram.

The process begins with the *Identification* phase, where initial searches from academic databases yielded 494 publications. An additional ten were found through other methods, specifically the snowballing method. Of these combined 504 publications, 22 duplicates were removed, leaving 482 unique publications.

The *Screening* stage involved two steps. First, titles and abstracts of the 482 publications were reviewed for relevance to the study. 392 unrelated publications were excluded. The remaining 90 publications underwent further screening through a partial review, during which a set of predefined inclusion and exclusion criteria was applied:

1. Publications were selected if:
   - The publication is peer-reviewed or part of recognized gray literature[10].
   - The publication must report the use of machine learning models for conflict forecasting.
2. Publications were excluded if:
   - The publication is not related to the intersection of conflict forecasting and machine learning.
   - The full text of the publication is not accessible.
   - The publication is not written in English.
   - The publication does not provide a DOI.

This filtering resulted in the exclusion of 58 articles, leaving 32 studies that satisfied all criteria. An overview table summarizing these publications is available at: Overview table.

2.3 Synthesis

The third phase of the SALSA strategy, synthesis, focuses on analyzing the selected publications to identify common patterns, trends, and gaps within the existing literature on ML applications in conflict forecasting.

To conduct a structured analysis of the 32 identified publications, a set of guiding sub-questions was developed, structured specifically to obtain detailed insights relevant to the research question. These questions were grouped into five thematic categories, each representing a critical dimension of how ML is applied in the context of conflict forecasting. The five themes were chosen to capture both the technical and practical considerations of ML use, particularly as they relate to anticipatory strategies in humanitarian settings.

1. *Forecasting scope and purpose*: Examines the geographic focus, the specific conflict-related events being predicted, and the intended use of the forecasts.
2. *Data sources and quality*: Investigates the datasets used to develop the models, with attention to the reported data issues.
3. *Modelling approaches*: Describes the employed model types, along with the applied validation techniques.
4. *Reliability and robustness*: Assesses how uncertainty is quantified and the extent of error analysis.
5. *Ethics and practical application*: Evaluates ethical considerations, including alignment with humanitarian principles, potential sources of bias, and whether the models are deployed in the real world or not.

The full set of sub-questions can be found in Appendix A. Each of the 32 selected publications was reviewed using these thematic questions.

2.4 Analysis

The final phase of the SALSA strategy, analysis, involves interpreting the synthesized findings to generate insights and relate the findings back to the research questions.

Specifically, the analysis sought to identify patterns and themes across the five thematic areas mentioned in Subsection 2.3. For each area, a cross-comparative evaluation was conducted to identify shared practices, differences, gaps, and unresolved challenges. This involved labeling responses to identify recurring approaches, similarities and differences across studies, and common challenges in methodology, data use, and application scope.

Particular attention was paid to the validation methods used, transparency in model reporting, and whether the proposed ML systems had documented real-world applications because these elements are critical indicators of a model's reliability and practical utility in humanitarian settings. In high-stakes environments, where forecasts inform decisions about resource allocation or early intervention, robust validation is essential to ensure predictive accuracy. Transparency in reporting creates trust among stakeholders, including humanitarian organizations and affected populations. Finally, documented real-world applications demonstrate feasibility and offer insights into how theoretical models perform under

[10] Gray literature refers here to non-peer-reviewed yet reputable sources such as reports, datasets, and working papers produced by humanitarian organizations and international agencies.

operational constraints such as limited data or ethical considerations. Evaluating these factors helps assess not only the technical quality of the models but also their readiness for real-world use in fragile contexts.

The findings derived from this analysis are presented in the Results chapter (Section 3).

3 Results

3.1 Forecasting Scope and Purpose

The geographic scope of the included studies varies notably, as illustrated in Figure 2. A large portion of the studies (14 out of 32) are aiming to develop models that provide forecasts across multiple countries. This includes systems such as the ViEWS project [21], the European Union Conflict Warning System [22], and the ACLED Conflict Alert System (CAST) [23]. Ten studies focus on Africa, particularly Sub-Saharan Africa, highlighting the region's importance due to persistent instability. The remaining studies are narrower in scope, targeting individual countries, including Bangladesh, Syria, Colombia, Indonesia, and Tunisia.

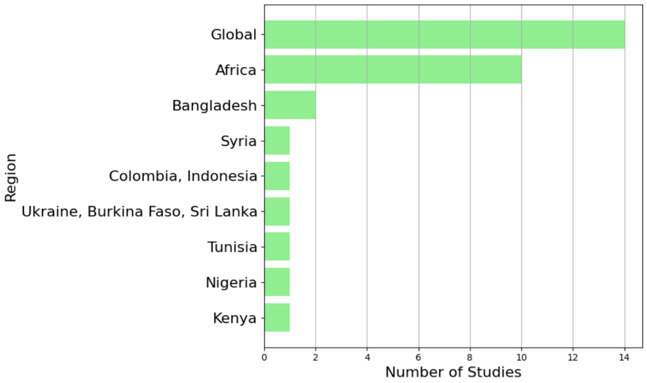

Figure 2: Geographic focus of included studies

In addition to geographic diversity, the included studies differ in terms of the specific conflict-related outcomes they aim to forecast. As shown in Figure 3, the most common target is the occurrence of conflict, with 14 out of 32 studies forecasting whether conflict will take place in a given unit of analysis, such as a country or subnational region, during a defined forecast window (e.g., monthly, yearly), depending on the study. Notably, one of these studies focuses specifically on the occurrence of terrorist events [24]. Following this, 12 studies aim to predict the onset of conflict, meaning the appearance of conflict after a period of peace. Among these, two studies specifically target the onset of civil wars. Nine studies estimate the number of future conflict fatalities, while six studies focus on predicting the type of conflict. One study forecasts the number of battles in a given region and time period [25]. Significantly, one study predicts the duration of peace agreements, estimating how long peace is likely to last before conflict recurs [26]. Lastly, one study extends the scope of terrorism forecasting by predicting not only the occurrence but also the location, type of attack, suspected group, and target involved in a future terrorist event [27].

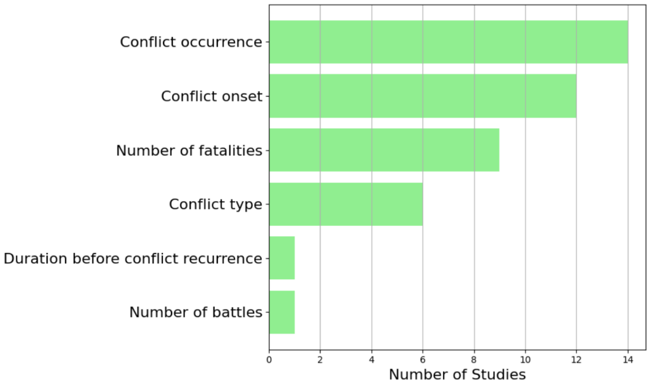

Figure 3: Forecasted outcomes focus of included studies

Next, the intended purposes identified by the studies fall into distinct categories:

- Early warning refers to forecasting conflict to enable timely alerts and preventive interventions by relevant actors, such as governments or peacekeeping organizations.
- Policy-making support involves generating forecasts that inform the development or adjustment of public policies.
- Risk assessment focuses on identifying areas or populations that are most vulnerable to conflict.
- Strategic planning is oriented toward long-term decision-making.
- Methodological improvement is concerned with advancing the technical aspects of forecasting, such as increasing accuracy or interpretability.

18 of the reviewed models position themselves as operational early warning systems designed to enable timely responses to anticipated conflict events. Another nine highlight their utility in informing policy-making processes. Methodological improvement is stated as an explicit goal, with ten studies focusing on improving methodological robustness rather than deploying operational systems.

Table 1 illustrates that 15 studies simultaneously serve multiple purposes, like ViEWS [21] and ACLED CAST [23], which extend their practical utility across policy-making, risk assessment, and strategic planning contexts.

3.2 Data Sources and Quality

The models included in this study rely on a wide range of data sources. The most frequently used source is the ACLED (Armed Conflict Location & Event Data) dataset, which appears in 13 out of the 32 reviewed studies. The Georeferenced Event Dataset (GED) from the Uppsala Conflict Data Program (UCDP) is also widely used, featuring in 10 studies. In addition to conflict event data, many models incorporate demographic and socioeconomic indicators sourced from the World Bank Open Data platform[11].

[11] https://data.worldbank.org/

Table 1: Forecasting goals identified across the reviewed studies. Some studies serve multiple purposes.

Forecasting goal	Studies
Early warning	[21], [22], [23], [25], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]
Policy-making support	[21], [22], [27], [28], [29], [32], [33], [41], [42]
Risk assessment	[23], [26], [35], [36], [40], [41], [43], [44]
Strategic planning	[21], [23], [24], [25], [29]
Methodological improvement	[26], [44], [45], [46], [47], [48], [49], [50], [51], [52]

Other commonly used datasets include the ViEWS dataset, the Civil War Dataset (CWD), the ICEWS (Integrated Crisis Early Warning System) dataset, the Power-Sharing Event Dataset (PSED), the Global Terrorism Database (GTD), the Rulers, Elections, and Irregular Governance (REIGN) dataset, and the V-Dem (Varieties of Democracy) dataset. These sources provide diverse coverage of political institutions, leadership, regime characteristics, and patterns of political violence.

Several studies also combine structured datasets with textual data from sources such as news articles, social media, and national surveys. Notably, one study applies the Latent Dirichlet Allocation (LDA) topic modelling technique to extract conflict-relevant signals from millions of newspaper articles [29]. Another study integrates data from a two-wave national survey to capture sociopolitical dynamics, such as perceived insecurity and trust in government, that are often absent from event-based datasets [46].

The ACLED dataset is widely utilized. However, in 2013, Perry cited concerns raised by Kristine Eck about quality-control issues in the ACLED dataset, particularly cautioning that "those interested in sub-national analyses of conflict should beware of ACLED's data due to quality-control issues which can result in biased findings if left unchecked by the researcher" [25]. Almost a decade later, Macis et al. [28] describe the ACLED dataset as "valuable in predicting conflict", emphasizing its utility for ML models due to its disaggregated structure and regular updates. They also highlight that the dataset is publicly accessible, making it well-suited for transparent and reproducible forecasting applications. Similar concerns regarding data quality have been raised by Blair and Sambanis [45] regarding the ICEWS dataset, which is recognized to occasionally suffer from noise and misclassification due to automated event coding processes.

Despite this diversity of datasets used, several common challenges appear regarding data quality, preprocessing, and bias management. A frequent issue is missing data, which can significantly affect the predictive accuracy of models if not addressed appropriately. Most studies handle missing data through various data imputation strategies, such as Multiple Imputation by Chained Equations (MICE), which models missing values multiple times; Last Observation Carried Forward (LOCF), which fills gaps with the last known value; and random forest imputation, which predicts missing values using decision trees.

Class imbalance presents another common issue, especially problematic due to the rarity of conflict events compared to non-conflict cases. Without addressing class imbalance, models risk trivial predictive success by mostly forecasting "no conflict," which is accurate but practically ineffective for conflict prevention or mitigation efforts. To manage this imbalance, studies typically employ strategies such as downsampling [44], which intentionally reduces the majority class cases to balance the dataset, or the Synthetic Minority Oversampling Technique (SMOTE) [31], which generates synthetic data points for minority classes.

Moreover, data preprocessing frequently involves steps such as the removal of duplicates and filling null values with placeholders. In some instances, studies specifically filter out zero-fatality events or redundant fields to enhance analytical clarity and model performance [24].

3.3 Modelling Approaches

The 32 reviewed studies employ a diverse range of ML techniques. As shown in Figure 4, Random Forest is the most frequently used algorithm, appearing in more than half of the studies. Logistic regression is also often used, either as the primary forecasting model or as a baseline in comparative evaluations. For example, Muchlinski et al. [44] implemented a Random Forest model and compared its performance with logistic regression. Although logistic regression is often categorized as a statistical method, models that use it are included here under the ML category, as they are implemented as supervised learning algorithms with automated prediction pipelines in all reviewed cases.

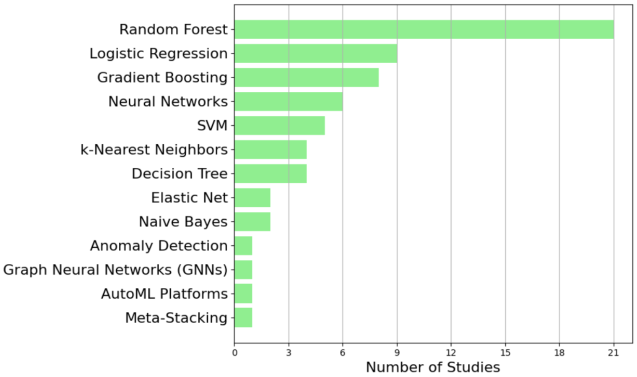

Figure 4: Frequency of the ML algorithms

Beyond these models, the studies implement gradient boosting methods, such as XGBoost, Support Vector Machines (SVMs), and various types of neural networks, including Artificial Neural Networks (ANNs), long short-term memory networks, and Graph Neural Networks (GNNs). A number of studies also explore novel or less conventional modelling strategies to enhance predictive power. For instance, AutoML frameworks were used to automate model selection and tuning [47]. Anomaly detection approaches using autoencoders were explored to identify unexpected shifts in violence patterns [37].

Model evaluation methods vary across the reviewed studies. The most common approach is cross-validation, particularly 10-fold cross-validation, which is used in 23 studies. In the case of Random Forest models, cross-validation is often combined with out-of-bag (OOB) error estimation, a built-in mechanism that provides nearly unbiased error estimates and helps prevent overfitting. Other studies use simpler fixed train/test splits to evaluate model performance. A few employ time-based splits for out-of-sample validation, such as training on data from 2001–2007 and testing on 2008–2015 in the case of Blair and Sambanis [45].

However, model evaluation practices are inconsistent. Some studies report detailed performance metrics, including Area Under the Curve (AUC), precision, recall, F1 score, and Brier scores, offering a detailed view of model effectiveness. In contrast, others mention the use of cross-validation but don't provide specific metrics or include confusion matrices, making it difficult to assess their predictive quality.

3.4 Reliability and Robustness

The studies included in this analysis show variation in terms of how they assess the robustness of their forecasts.

Uncertainty in the included studies is mostly addressed indirectly through the performance evaluation of ML models. As described in Subsection 3.3 most studies rely on out-of-sample evaluations and cross-validation methods to assess and validate the robustness of their models. These methods mitigate model uncertainty by repeatedly testing the model on unseen data.

Several classification models frequently employ standard metrics such as Area Under the ROC Curve (AUC-ROC) and Area Under the Precision-Recall Curve (AUC-PR) to quantify prediction uncertainty. Similarly, regression-oriented studies commonly use Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as performance measures. However, explicit quantification of model uncertainty, such as confidence or prediction intervals, is rare. Only two studies leverage bootstrapped confidence intervals or conduct simulation-based analyses to characterize model robustness [43; 51].

Regarding error analysis, studies predominantly rely on aggregate quantitative metrics. While standard performance indicators such as AUC-ROC, accuracy, and precision-recall curves are reported, detailed error analyses are rare. Only eight studies provide more thorough insights, including confusion matrices, false-positive and false-negative breakdowns, and region-specific or temporal evaluations. Comprehensive analyses, often seen in operational systems like ViEWS [21] or ACLED CAST [23], are exceptions rather than the norm.

3.5 Ethics and Practical Application

Among the 32 studies reviewed, only five mention or imply ethical considerations. These few studies that do address ethics tend to focus on themes such as interpretability, transparency, media bias, uncertainty communication, and risks of misuse. For instance, one study emphasizes the importance of model interpretability to ensure that predictions are transparent and actionable for policymakers [28]. Another study mentions ethical concerns tied to the use of predictive systems in counter-terrorism, including potential issues related to surveillance, wrongful profiling, and fairness [24]. Transparency is often highlighted as an ethical value, with some authors not only acknowledging biases, such as those stemming from media censorship, but also taking steps to mitigate them by publishing forecasts, code, and datasets openly and ensuring regular updates [29]. In addition, the communication of uncertainty is framed as an ethical imperative, particularly when predictions are directed at non-expert stakeholders like humanitarian actors or policy officials [43]. Perhaps most notably, the ViEWS project explicitly warns of the possibility that its forecasts could be misused, for example, by enabling governments to justify preemptive violence [21]. To mitigate this risk, the project embraces transparency and public accessibility, positioning its platform as a tool to support civil society, NGOs, and international organizations in conflict prevention.

Additionally, despite the increasing sophistication of ML models, only five of the reviewed systems are currently in operational use. The ACLED CAST delivers monthly updated forecasts through a public dashboard, providing a practical interface for early warning [23]. The Global Conflict Risk Index is another example of applied forecasting, serving as the quantitative foundation for the European Union's Conflict Early Warning System [22]. The model introduced by Mueller et al. [29] is made available through the Conflict Forecast[12] platform, which is used not only in research but also in real-world policy settings, including the UK's Foreign, Commonwealth & Development Office (FCDO) intervention in Nigeria. Lastly, the ViEWS project, active since 2018, provides monthly forecasts of political violence and is publicly available for use by NGOs, researchers, and international institutions [21].

4 Discussion

4.1 Survey Findings

Firstly, out of the 32 studies included in this survey, 27 are published after 2020. This indicates a growing research interest in using ML for conflict forecasting, particularly in the time of global instability. In terms of geographic focus, 10 of the reviewed studies target Africa, highlighting the region's relevance due to persistent instability and humanitarian risk. However, the majority of studies are focusing on a global approach rather than targeting a specific region. This fact may enhance model generalizability but comes with the risk of overlooking local contextual factors that are important for practical applications. Additionally, the results

[12] https://conflictforecast.org/

show that most of the studies have as their intended purpose early warning, but also another considerable number aim for methodological improvement, as can be seen in Table 1. This division in goals reflects the dual focus of the field: while some researchers prioritize building models that can be used for operational forecasting, others concentrate on advancing the technical aspects of conflict prediction through academic experimentation. Moreover, the surveyed studies vary significantly in their forecasting targets. This diversity in goals means that studies are not always directly comparable, complicating broad generalizations. This underscores the importance of context-specific evaluation: insights into the feasibility of ML models are better understood within the specific framing of each study rather than generalized across the entire field.

Secondly, the analysis demonstrates that a wide variety of datasets are used, with ACLED and UCDP being the most common. Many studies identify issues with missing, biased, or inconsistent data, especially for the most conflict-vulnerable regions. Areas such as Sub-Saharan Africa often suffer from limited reporting capacity, which reduces the trustworthiness of the data used to train ML models. Although the studies propose mitigation strategies, as described in Subsection 3.2, these methods may not fully resolve the core problem. Consequently, the practical accuracy of these forecasts in underreported regions remains questionable, creating risks to humanitarian organizations that might rely on them. An unreliable forecast could lead to misallocation of already insufficient resources or leave at-risk populations unsupported.

Thirdly, the multitude of different modelling approaches reflects the complexity of conflict prediction. As discussed in Subsection 3.3, the most frequently used algorithm is Random Forest, which is widely recognized for performing well in high-dimensional settings and for its ability to capture nonlinear relationships between variables. As Biau and Scornet [53] highlight, this algorithm is well-suited to settings involving numerous predictors, which is a condition met by many conflict forecasting tasks. For instance, Rød et al. [51] combine Random Forest with a diverse set of variables from ACLED, V-Dem, REIGN, and the World Bank data. Furthermore, the results show that model validation practices vary considerably. Most studies implement cross-validation, but the transparency and granularity of performance reporting differ. This lack of consistency makes it difficult to compare models and may hide issues such as overfitting or differences in performance across regions or time periods.

Fourth, an important limitation found across the survey is the inconsistent reporting of uncertainty and error analysis. Although some papers acknowledge the high-stakes nature of the humanitarian decisions these models aim to support, only a minority quantify prediction uncertainty or explore sources of error. This highlights potential areas for further standardization and improvement within conflict forecasting research. In this domain, uncertainty needs to be clearly specified for having accurate and trustworthy forecasts, and if the reporting of uncertainty and the error analysis are not standardized, it is hard to conduct a good comparison between the studies.

Finally, the review found that only five out of 32 models mention ethical implications of deploying ML in humanitarian settings. Additionally, only five studies report practical deployment, while the remaining models primarily serve as proof-of-concept or methodological exploration. This indicates a significant gap between academic modelling and operational use. While the lack of deployment may reflect institutional barriers or data limitations, it also raises concerns about whether the research community is sufficiently addressing the translational challenges of moving from model development to field implementation. Notably, out of the five deployed models, only two engage with ethical considerations. As humanitarian decisions often involve vulnerable populations and limited budgets, overlooking ethical implications or deployment constraints may inadvertently cause harm or undermine stakeholder trust in these systems.

4.2 Answering the Research Questions

Building on the patterns and gaps identified in Subsection 4.1, this subsection interprets the findings in order to address the two sub-questions and, ultimately, the main research question.

Trustworthiness and Accuracy of ML Models (RQ1)

This review finds that although many models report promising predictive accuracy, their overall trustworthiness is limited by inconsistent evaluation and reporting practices. In particular, few studies assess the implications of false positives and false negatives, an omission with serious consequences in humanitarian contexts. False negatives may lead to missed early interventions, while false positives can result in the misallocation of scarce resources. Only a few papers provide confusion matrices or performance breakdowns that could support such analysis. These gaps point to a broader lack of standardization in validation protocols within this field. As Agbabiaka et al. [54] emphasize in their review of trustworthy AI in public-sector decision-making, trustworthiness demands regular, transparent evaluation and communication of model performance. Without this, even technically sophisticated models cannot be considered reliable or actionable in operational humanitarian settings.

Practical Feasibility of Model Deployment (RQ2)

The practical deployment of conflict forecasting models in humanitarian settings depends on several contextual conditions that are often overlooked. First, many conflict-prone regions lack reliable, timely, and disaggregated data. In such cases, model outputs risk being misleading, which can result in harmful decisions.

Second, operational feasibility remains a barrier. Most organizations in the field face a limited budget. For a model to be useful, it must be both cost-effective and interpretable. Highly complex systems, even if technically sound, are unlikely to be adopted unless their outputs are justifiable to non-technical decision-makers.

Third, few models communicate uncertainty in ways that are actionable. In real-world forecasting, uncertainty is essential for decision-making. Yet most reviewed studies fail to report uncertainty, limiting their value for planning under risk.

Finally, ethical concerns must be taken seriously. Without safeguards, models may reinforce surveillance biases or be misused to justify political agendas. Transparency, accountability, and clear documentation are needed to prevent such outcomes.

In short, forecasting models can only support humanitarian decisions when their outputs are data-informed, interpretable, uncertainty-aware, and ethically grounded. Without these conditions, deployment remains limited, regardless of technical performance.

Addressing the Main Research Question

Taken together, the findings suggest that while ML systems for conflict forecasting show technical potential, they are not yet sufficiently reliable or feasible for widespread use in real-world humanitarian contexts. Their trustworthiness is limited by inconsistent evaluation practices, lack of standardized error analysis, and poor communication of uncertainty. At the same time, their practical deployment is constrained by low interpretability and a general absence of ethical and operational integration. As a result, most models remain confined to academic experimentation, with only a few transitioning into real-world applications. To bridge this gap, future work in the field must prioritize not just accuracy, but also usability, transparency, and ethical safeguards. Only then can ML-based forecasting systems meaningfully support humanitarian decision-making.

5 Responsible Research

This study adheres to the five principles of responsible research conduct, according to the Netherlands Code of Conduct for Research Integrity [55]. These five principles are honesty (truthful representation of methods and findings), scrupulousness (careful and precise execution of research), transparency (open communication about research processes and outcomes), independence (freedom from outside influence or conflicts of interest), and responsibility (awareness of the societal impact of research). The application of these principles is reflected in the following aspects of the research process:

- Honesty has guided the literature review through accurate representation of sources, faithful interpretation of authors' arguments, and avoidance of misquotation or selective reporting.
- Scrupulousness has been upheld by consistently using citation standards and maintaining precision in analyzing, citing, and comparing the literature.
- Transparency has been ensured by clearly describing the criteria for the selection of the articles and comparison methods. The systematic literature review process ensures the reproducibility and allows others to understand and verify the research approach.
- Independence has been respected by forming judgments and perspectives based on critical engagement with the literature, free from personal, institutional, or ideological bias.
- Responsibility has been demonstrated by acknowledging the ethical relevance and societal implications of the research domain, and by engaging with the literature in a respectful and constructive manner.

In addition, a few tools were used to improve the readability of the paper. Grammarly[13] was used for grammar checking. ChatGPT[14] and QuillBot[15] were used to assist with rephrasing during the writing process. As English is my second language, I occasionally find it challenging to express ideas in academic style, and these tools helped me refine the language while maintaining the originality of the content. The prompt used for ChatGPT was: *"Help me rephrase this paragraph to make it sound more academic and natural. I want the tone to be appropriate for a research paper."*.

6 Limitations and Future Work

6.1 Limitations

This study is subject to several limitations that may influence its findings. Firstly, the project was conducted over a period of ten weeks, which constrained the depth of the literature search, synthesis, and analysis. While care was taken to select the relevant sources, the time constraints may have limited the inclusion of other valuable studies.

Secondly, the review was restricted to papers that are publicly available or available through databases that are accessible to TU Delft. As a result, relevant studies published in subscription-only journals may have been excluded, potentially affecting the completeness of the review.

Lastly, only English-language literature was considered. This introduces a language bias, as models published in other languages were not captured. This may have led to the omission of culturally important insights.

6.2 Future Work

This review primarily focused on peer-reviewed academic literature, with the exception of one gray literature source: the ACLED CAST [23]. Obtaining detailed documentation on the methodologies used by the humanitarian organizations in their forecasting models was not feasible within the project's timeframe. As a result, the review relied more heavily on sources with publicly available and well-documented methods, which led to a stronger focus on academic research.

Future work should aim to systematically include gray literature, such as internal reports and model documentation from humanitarian agencies and NGOs. These sources often offer crucial insights into real-world implementation, model adaptation, and region-specific challenges that are underrepresented in academic studies. Additionally, expanding the review to include non-English publications would help capture more diverse perspectives, especially from conflict-affected regions where relevant work may not be published in English.

Another important direction for future work is a more in-depth examination of the datasets used in conflict forecasting models. Analyzing the reasoning behind dataset selection,

[13] https://www.grammarly.com/

[14] https://chatgpt.com/

[15] https://quillbot.com/

comparing available alternatives, and evaluating trade-offs in coverage, granularity, and update frequency would contribute to a better understanding of how data choices influence model performance and relevance across different operational settings.

For researchers developing ML models for conflict forecasting, we also recommend the adoption of systematic reporting practices for model uncertainty, error analysis, and validation methods. Clear documentation of these aspects is essential for assessing model robustness, ensuring transparency, and supporting responsible deployment in high-stakes humanitarian contexts.

7 Conclusion

This study investigated whether ML models for conflict forecasting are reliable and feasible for application in humanitarian contexts. It evaluated the strengths, limitations, and potential applications of these models, focusing on their relevance for humanitarian decision-making. The assessment addressed the models' geographic scope, the kinds of outcomes they forecast, and the forecasting purposes they serve. It also examined how the models are validated, the types of data they use, and the model types they employ. Finally, it evaluated how they handle uncertainty, report errors, consider ethical implications, and whether the models are used in practice.

The review found that while ML-based conflict forecasting models show promise, their practical utility in humanitarian settings remains limited. Most models lack standardization in performance validation and do not systematically address key aspects such as uncertainty quantification or ethical implications. Only a few are deployed in operational settings, and many remain confined to academic experimentation. Common challenges include inconsistent reporting, insufficient interpretability, and unreliable or missing data, particularly in regions most affected by conflict. These factors constrain both the trustworthiness and the real-world applicability of such models, creating risks for humanitarian decision-making.

Future research should prioritize integrating gray literature from humanitarian organizations to better understand the operational use of ML models developed by humanitarian organizations. Expanding the scope to include non-English sources would also help surface insights from conflict-affected regions often excluded from academic discourse. Another important direction for future work is a more in-depth examination of the datasets used in conflict forecasting models. Only by addressing these gaps can ML-based forecasting systems move beyond academic potential and meaningfully support humanitarian action.

A List of sub-questions for Synthesis

1. *Forecasting scope and purpose:*
   - What is being forecasted?
   - For what purpose is it being forecasted?
   - What is the geographical coverage?
2. *Data sources and quality:*
   - What data is used?
   - How are data quality issues and biases handled?
3. *Modelling approaches:*
   - What type of model is used?
   - How is the model validated?
4. *Reliability and robustness:*
   - How is uncertainty addressed?
   - How thorough is the error analysis?
5. *Ethics and practical application:*
   - What ethical considerations are discussed?
   - Is the model used in practice?

References

[1] Håvard Hegre, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal. Predicting Armed Conflict, 2010–2050. *International Studies Quarterly*, 57(2):250–270, June 2013.

[2] International Committee of the Red Cross. ICRC in 2024: Upholding Humanity in Conflict, December 2024. Accessed: 2025-05-23.

[3] Council on Foreign Relations. Ukraine: Conflict at the Crossroads of Europe and Russia, February 2023. Accessed: 2025-05-23.

[4] Shawn Davies, Therése Pettersson, and Magnus Öberg. Organized violence 1989–2022, and the return of conflict between states. *Journal of Peace Research*, 60(4):691–708, 2023.

[5] Omar Shahabudin McDoom. Expert Commentary, the Israeli-Palestinian Conflict, and the Question of Genocide: Prosemitic Bias within a Scholarly Community? *Journal of Genocide Research*, pages 1–9, 2024.

[6] Development Initiatives. Falling Short? Humanitarian Funding and Reform, 2024. Accessed: 2025-05-23.

[7] Aryn Baker. What Man, and Climate Change, Has Wrought. *TIME*, March 2017. Accessed: 2025-05-23.

[8] High-Level Panel on Humanitarian Financing. Too Important to Fail: Addressing the Humanitarian Financing Gap, 2016. Accessed: 2025-05-23.

[9] Observer Research Foundation. The Growing Gaps in Global Humanitarian Challenges, 2024. Accessed: 2025-05-23.

[10] Olivia O'Sullivan and Jerome Puri. First USAID closes, then UK cuts aid: what a Western retreat from foreign aid could mean, March 2025. Accessed: 2025-05-23.

[11] Erin C. Lentz and Daniel Maxwell. How Do Information Problems Constrain Anticipating, Mitigating, and Responding to Crises? 81:103242.

[12] World Bank and United Nations. *Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict*. World Bank, Washington, DC, September 2017.

[13] Lars-Erik Cederman and Nils B. Weidmann. Predicting Armed Conflict: Time to Adjust Our Expectations? *Science*, 355(6324):474–476, 2017.

[14] Tina Comes. AI for Crisis Decisions. *Ethics and Information Technology*, 26(1):12, February 2024.

[15] David Paulus, Gerdien de Vries, Marijn Janssen, and Bartel Van de Walle. Reinforcing Data Bias in Crisis Information Management: The Case of the Yemen Humanitarian Response. 72:102663.

[16] Timur Obukhov and Maria A. Brovelli. Identifying Conditioning Factors and Predictors of Conflict Likelihood for Machine Learning Models: A Literature Review. *ISPRS International Journal of Geo-Information*, 12(8), 2023.

[17] Espen Geelmuyden Rød, Tim Gåsste, and Håvard Hegre. A review and comparison of conflict early warning systems. *International Journal of Forecasting*, 40(1):96–112, 2024.

[18] Maria J. Grant and Andrew Booth. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2):91–108, 2009.

[19] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, New York, NY, USA, 2014. Association for Computing Machinery.

[20] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[21] Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby. Views: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174, 2019.

[22] Matina Halkia, Stefano Ferri, Marie K. Schellens, Michail Papazoglou, and Dimitrios Thomakos. The Global Conflict Risk Index: A quantitative tool for policy support on conflict prevention. *Progress in Disaster Science*, 6:100069, 2020.

[23] ACLED. ACLED Conflict Alert System. https://acleddata.com/conflict-alert-system/, July 2023. Accessed 2025-06-10.

[24] Tosin Comfort Olayinka, Lawal Adeniyi, Duke Oghorodi, Kizito Eluemunor Anazia, Ojei Harrison Onyijen, Chukwuemeka Pascal Nwankwo, Akinola Samson Olayinka, Wilson Nwankwo, and Kingsley Ukaoha. Patterns of Terror: A Comparative Predictive Model. In *2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON)*, pages 1–5, 2024.

[25] Chris Perry. Machine Learning and Conflict Prediction: A Use Case. *Stability: International Journal of Security amp; Development*, 2(3):56, October 2013.

[26] Andrew B. Whetten, John R. Stevens, and Damon Cann. The implementation of random survival forests in conflict management data: An examination of power sharing and third party mediation in post-conflict countries. *PLOS ONE*, 16(5):e0250963, May 2021.

[27] Asma El Kissi Ghalleb and Najoua Essoukri Ben Amara. Terrorist Act Prediction Based on Machine Learning: Case Study of Tunisia. In *2020 17th International Multi-Conference on Systems, Signals Devices (SSD)*, pages 398–403, 2020.

[28] Luca Macis, Marco Tagliapietra, Alessandro Castelnovo, Daniele Regoli, Greta Greco, Andrea Claudio Cosentini, Paola Pisano, and Edoardo Carroccetto. Integrating XAI for Predictive Conflict Analytics. In *Proceedings of the 2nd World Conference on eXplainable Artificial Intelligence (XAI)*. CEUR Workshop Proceedings, 2024. Late-breaking work, Demos and Doctoral Consortium, Valletta, Malta, July 17–19.

[29] Hannes Mueller, Christopher Rauh, and Ben Seimon. Introducing a global dataset on conflict forecasts and news topics. *Data 38; Policy*, 6:e17, 2024.

[30] Hannes Mueller and Christopher Rauh. Using past violence and current news to predict changes in violence. *International Interactions*, 48(4):579–596, May 2022.

[31] Mark Musumba, Naureen Fatema, and Shahriar Kibriya. Prevention Is Better Than Cure: Machine Learning Approach to Conflict Prediction in Sub-Saharan Africa. *Sustainability*, 13(13):7366, July 2021.

[32] H. M. Mahmudul Hasan, Adil Ahnaf, and Nahid Hossain. Prediction of Political and Local Conflicts in Bangladesh: An Event Analysis. In *2021 International Conference on Science Contemporary Technologies (ICSCT)*, pages 1–6, 2021.

[33] Çağlar Akar, Doğa Başar Sarıipek, and Gökçe Cerev. Poverty-Armed Conflict Nexus: Can Multidimensional Poverty Data Forecast Intrastate Armed Conflicts? *Social Inclusion*, 12, September 2024.

[34] Lungisani Ndlovu, Nenekazi Mkuzangwe, Anton De Kock, Ntombizodwa Thwala, Japhtalina Mokoena, and Rethabile Matimatjatji. A Situational Awareness Tool using Open-Source Intelligence (OSINT) and Artificial Intelligence (AI). In *2023 IEEE International Conference on Advances in Data-Driven Analytics And Intelligent Systems (ADACIS)*, pages 1–6, 2023.

[35] Felix Ettensperger. Forecasting conflict using a diverse machine-learning ensemble: Ensemble averaging

with multiple tree-based algorithms and variance promoting data configurations. *International Interactions*, 48(4):555–578, December 2021.

[36] Quansheng Ge, Mengmeng Hao, Fangyu Ding, Dong Jiang, Jürgen Scheffran, David Helman, and Tobias Ide. Modelling armed conflict risk under climate change with machine learning and time-series data. *Nature Communications*, 13(1), May 2022.

[37] Luca Macis, Marco Tagliapietra, Rosa Meo, and Paola Pisano. Breaking the trend: Anomaly detection models for early warning of socio-political unrest. *Technological Forecasting and Social Change*, 206:123495, September 2024.

[38] Samuel Bazzi, Robert A. Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Richard Peck. The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia. *The Review of Economics and Statistics*, 104(4):764–779, July 2022.

[39] Daniel Racek, Paul W. Thurner, Brittany I. Davidson, Xiao Xiang Zhu, and Göran Kauermann. Conflict forecasting using remote sensing data: An application to the Syrian civil war. *International Journal of Forecasting*, 40(1):373–391, January 2024.

[40] Nicholas Shallcross and Darryl Ahner. Predictive models of world conflict: accounting for regional and conflict-state differences. *The Journal of Defense Modeling Simulation*, 17:243–267, July 2019.

[41] Sondip Poul Singha, Md. Mamun Hossain, Md. Ashiqur Rahman, and Nusrat Sharmin. Investigation of graph-based clustering approaches along with graph neural networks for modeling armed conflict in Bangladesh. *International Journal of Data Science and Analytics*, 18(2):187–203, June 2024.

[42] Jannis M. Hoch, Sophie P. de Bruin, Halvard Buhaug, Nina Von Uexkull, Rens van Beek, and Niko Wanders. Projecting armed conflict risk in Africa towards 2050 along the SSP-RCP scenarios: a machine learning approach. *Environmental Research Letters*, 16(12):124068, December 2021.

[43] David Randahl, Jonathan Williams, and Håvard Hegre. Bin-Conditional Conformal Prediction of Fatalities from Armed Conflict, October 2024.

[44] David Muchlinski, David Siroky, Jingrui He, and Matthew Kocher. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1):87–103, 2016.

[45] Robert A. Blair and Nicholas Sambanis. Forecasting Civil Wars: Theory and Structure in an Age of "Big Data" and Machine Learning. *Journal of Conflict Resolution*, 64(10):1885–1915, April 2020.

[46] Andrew M. Linke, Frank D.W. Witmer, and John O'Loughlin. Weather variability and conflict forecasts: Dynamic human-environment interactions in Kenya. *Political Geography*, 92:102489, January 2022.

[47] Vito D'Orazio, James Honaker, Raman Prasady, and Michael Shoemate. Modeling and Forecasting Armed Conflict: AutoML with Human-Guided Machine Learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4714–4723, 2019.

[48] Jonathan Pinckney and Babak RezaeeDaryakenari. When the levee breaks: A forecasting model of violent and nonviolent dissent. *International Interactions*, 48(5):997–1026, August 2022.

[49] Fulvio Attinà, Marcello Carammia, and Stefano M. Iacus. Forecasting change in conflict fatalities with dynamic elastic net. *International Interactions*, 48(4):649–677, July 2022.

[50] Benjamin J. Radford. High resolution conflict forecasting with spatial convolutions and long short-term memory. *International Interactions*, 48(4):739–758, March 2022.

[51] Espen Geelmuyden Rød, Håvard Hegre, and Maxine Leis. Predicting armed conflict using protest data. *Journal of Peace Research*, 62(1):3–20, September 2023.

[52] Iris Malone. Recurrent neural networks for conflict forecasting. *International Interactions*, 48(4):614–632, January 2022.

[53] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2):197–227, April 2016.

[54] Olusegun Agbabiaka, Adegboyega Ojo, and Niall Connolly. Requirements for trustworthy AI-enabled automated decision-making in the public sector: A systematic review. *Technological Forecasting and Social Change*, 215:124076, 2025.

[55] Netherlands Universities. Netherlands Code of Conduct for Research Integrity, 2018. Accessed: 2025-06-02.