UNIVERSITEIT VAN AMSTERDAM

VU VRIJE UNIVERSITEIT AMSTERDAM

MSc Computer Science

Software Engineering and Green IT

Master Thesis

Beyond Closed Doors: An Open-Source AI Framework for Forecasting Armed Conflict

by

Oleksandr Zakotianskyi

2790533

Supervisors:	dr. Anna Bon prof. dr. Julia Schaumburg
Date:	August 21, 2025
Credits & Period:	30 ECTS, January 2024 - August 2025

Abstract

Armed conflicts are increasing in frequency and severity worldwide, resulting in devastating humanitarian and economic consequences. Despite advances in conflict forecasting, leading academic early warning models often lack complete openness, limiting reproducibility and the potential for advancements in the field. Addressing this critical issue, this thesis presents an entirely open-source framework designed to reproduce and extend state-of-the-art early conflict forecasting models using publicly available datasets.

Specifically, we systematically evaluate three advanced machine learning approaches—XGBoost, AutoGluon, and TabPFN—across existing datasets from ViEWS and Conflict Forecast, as well as our own extended dataset based on the Uppsala Conflict Data Program (UCDP), covering a period from 1989 to 2024. Our results demonstrate that fully open-source models can match, and even slightly surpass, the predictive performance of current closed-source benchmarks, as measured by Precision-Recall AUC.

By openly sharing all developed pipelines, datasets, models, and evaluation frameworks developed in this thesis, we aim to foster transparency, reproducibility, and further collaborative innovation in early conflict forecasting research.

Key words: Early Conflict Forecasting, Civil War, AI for Good.

[2-4]

Contents

Abstract **i**

Nomenclature **iv**

1 Introduction **1**

2 Literature Review **3**

3 Datasets **7**
3.1 ViEWS dataset 7
3.2 Conflict Forecast dataset 8
3.3 Pipeline for UCDP dataset 8

4 Methodology **9**
4.1 Definition of the dependent variable 9
4.2 Definition of hard onsets 10
4.3 Models 10
4.3.1 XGBoost 10
4.3.2 AutoGluon 11
4.3.3 AutoGluon Hyperparameter Optimization 12
4.3.4 TabPFN 13
4.4 Model training approach 14
4.4.1 Yearly training scheme 15
4.4.2 Monthly training scheme 16
4.4.3 Optimal validation set size 18

5 Modeling Enhancements **20**
5.1 Stacking 20
5.2 Ensembling 21
5.3 Combining Stacking and Ensembling 21
5.4 News-feature experiments 22

6 Results **25**
6.1 Model Performance on ViEWS data 25
6.2 Model Performance on Conflict Forecast Data 26
6.3 Model Performance on UCDP data 28
6.4 Predictions into the real future 29

7 Discussion **31**
7.1 Ethical Considerations 32

8 Conclusion & Future Work **33**

References **35**

A UCDP Pipeline Implementation details **38**

B ROC-AUC Scores for all models **41**

B.1 ViEWS Dataset ROC-AUC Scores 41

B.2 Conflict Forecast Dataset ROC-AUC Scores 42

B.3 UCDP Dataset ROC-AUC Scores 42

Nomenclature

Abbreviations

Abbreviation	Definition
ECW	Early Conflict Warning
ECWS	Early Conflict Warning System
ViEWS	Violence & Impacts Early-Warning System
UCDP	Uppsala Conflict Data Program
CF	Conflict Forecat Research Group

1

Introduction

In recent years, armed conflicts have become increasingly common worldwide, reaching levels not seen for decades. In 2023 alone, there were 59 state-based conflicts globally, marking the highest number recorded since systematic tracking began in 1946 [1]. Even more troubling, the overall incidence of political violence has nearly doubled in just five years: from about 100,000 recorded conflict events in 2020 to roughly 200,000 in 2024 [2], [3]. Today, one in every eight people worldwide lives in an area directly affected by organised violence [3]. The human cost of these conflicts has been severe. Recent wars have caused hundreds of thousands of deaths across multiple regions. In 2024 alone, over 233,000 lives were lost due to violence, marking a significant 30% rise from the previous year [1]. Violent conflicts are clearly an urgent global challenge, highlighting significant gaps in our ability to predict and respond effectively.

Beyond devastating humanitarian losses, conflicts also impose massive economic burdens, creating long-term impacts on affected regions. Recent research highlights that timely conflict prevention yields substantial economic benefits. Investing in early preventive measures can generate enormous returns, with estimates ranging from $26 to $75 for every dollar spent in countries without recent violence. For countries experiencing recent conflicts, these returns could be as high as $103 per dollar spent [4]. Clearly, improving our ability to anticipate and mitigate conflicts is crucial, not only to save lives but also to secure economic stability and sustainable development.

The rapid rise in conflict frequency and its unpredictability underscore the pressing need for reliable early warning systems. Policymakers and international organisations increasingly recognise the importance of data-driven tools capable of identifying conflicts before they escalate. Significant progress has been made through advancements in political risk analysis, machine learning, and extensive use of large datasets. However, current early-warning models still face considerable challenges, particularly regarding reproducibility and transparency. For instance, models like ViEWS demonstrate strong performance yet provide limited access to their data. Similarly, Conflict Forecast offers partially closed-source data and methods, hindering independent verification and improvement efforts.

The primary objective of this thesis is to address existing limitations related to transparency and reproducibility in existing early conflict warning systems. We aim to answer the following research questions: (1) Can a fully open-source early conflict warning system achieve predictive performance comparable to current state-of-the-art systems such as those developed by ViEWS and Conflict Forecast? (2) How effectively can advanced open-source machine learning frameworks (XGBoost, AutoGluon, and TabPFN) perform in predicting conflict, especially given limitations in publicly available datasets? and (3) To what extent can stacking and ensembling techniques further enhance the predictive accuracy of early conflict forecasting models? By addressing these questions, this research seeks to demonstrate the feasibility and benefits of transparent, reproducible, and high-performing conflict forecasting systems.

Throughout the study, we systematically evaluate state-of-the-art machine learning methods on datasets publicly available datasets published by ViEWS and Conflict Forecast, as well as on our own dataset which we built to overcome limitations present in the available to us datasets. By achieving comparable and even slightly improved predictive accuracy (measured by Precision-Recall AUC), we demonstrate that open-source models can match or surpass current closed-source benchmarks. Furthermore, we provide our entire pipeline and framework openly, enabling future research and fostering collaboration[1].

This thesis is structured as follows: Section 2 reviews existing literature on early conflict forecasting. Section 3 describes the datasets used in our analysis. Section 4 outlines our methodology, including predictive variables and our training approach. Section 5 presents enhancements to our modelling approach, such as stacking and ensembling. Section 6 details our results, comparing our model performance with existing benchmarks. Section 7 discusses the implications of our findings, and Section 8 concludes by summarising contributions and suggesting avenues for future research.

In today's world, predicting and preventing violent conflict is not merely an academic question — it is a critical necessity.

[1] The code is available at `https://github.com/fif911/probabilistic_conflict_modelling`

2

Literature Review

Conflict early-warning systems have evolved over several decades, with their modern foundations laid in the 1980s. Early approaches focused on structural risk indicators – static country characteristics correlated with conflict – to identify states at risk of instability. These structural models treated conflict propensity as a function of underlying conditions (e.g. weak governance, economic grievances, demographic pressures) gleaned from historical data [5]. Governments and international organizations in the late Cold War era began investing in such predictive frameworks to anticipate crises [6]. For example, the State Failure Task Force (later PITF) in the 1990s developed country-year risk assessments using a fixed set of indicators (on political regime, societal fractionalization, economics, etc.) to forecast civil conflict and regime collapse [6], [7], [8]. These first-generation models were essentially logistic regressions or index scores built on past conflict patterns, and they provided valuable insight into long-term vulnerability but often missed the triggers of when violence would break out.

Over time, researchers sought to incorporate more dynamic warning signals beyond static structural data. A landmark contribution was the use of news media signals as predictors of conflict. Chadefaux demonstrated that patterns in news reporting can foreshadow interstate wars up to about one week in advance chadefaux2014early. Surges in conflict-related news coverage were shown to act as early warning signals, improving short-term prediction accuracy for war onsets. The growing availability of digital news feeds and event data in the 2000s thus opened the door for including real-time or high-frequency event indicators into early-warning models, complementing the slower-moving structural risk factors.

In the past decade, there has also been a pronounced turn toward machine learning techniques in this field. Advances in computing power and new datasets have enabled researchers to apply more advanced ML algorithms to conflict prediction tasks [9], [10]. Such models can automatically discover complex patterns and interactions in large conflict datasets, from country-year covariates to daily media reports. Indeed, recent reviews note that techniques from artificial intelligence, including ML and natural language processing are increasingly shaping modern conflict early-warning systems [6].

The result is a shift from purely theory-driven to more data-driven models. One of the most prominent data-driven efforts is the ViEWS project (Violence Early Warning System) based at Uppsala University [11], [12]. ViEWS provides systematic forecasts of armed conflict using both country-level and subnational data, updated on a monthly schedule. The project produces probabilistic early warnings for multiple forms of political violence – including state-based conflicts, non-state fighting, and one-sided violence against civilians – up to 36 months into the future.

A key innovation of ViEWS is its use of an ensemble of models built on diverse feature sets (conflict history, socio-economic indicators, political events, etc.), which are combined as tabular features to generate more robust predictions [11]. Notably, ViEWS integrates high-resolution data: it forecasts conflict not only at the national level but also across a grid of local regions (with an initial focus on Africa), thereby capturing subnational variation in conflict risk [12]. While, ViEWS published most of the code and documentation publicly, some models are not open-source, and on all full dataset are available for public use.

As one of the great initiatives, ViEWS invites external experts to contribute. In 2022 it organized the first prediction challenge, and in 2023 the second prediction competition [13], [14].

Another influential group pushing the frontier of early warning is the Conflict Forecast project led by Mueller and Rauh. This academic initiative applies machine learning and text analysis at a global scale to predict conflict, with a special focus on the "hard problem" of anticipating conflict onset in previously peaceful countries [15]. The Conflict Forecast uses two types of data: structural data (such as development indicators, governance measures, and recent violence history) and text news which are used to create news topic features from [16].

In practice, Conflict Forecast processes millions of news articles and derives 15 topics on the news background in each country using unsupervised topic modeling (Latent Dirichlet Allocation) [16], [17]. These textual features are combined with traditional predictors in an ensemble machine-learning framework, yielding monthly risk estimates for each country and PRIO grid cells [16].

While Conflict Forecast has previously released input datasets and model's code for public use and since 2022, no new reproducible code has been shared. [16]. In 2024 Conflict Forecast only shared the reduced dataset 2010 onwards, which contains only 40% of the data used to train their models without any code to reproduce the results. This hinders the reproducibility of their results and makes it difficult to improve over their existing models for researcher to advance the field.

Another notable development is a domain-specific large language model to analyze conflict-related text. ConfliBERT developed by Brandt et al. is a transformer-based language model pre-trained on a specialized corpus of conflict and political violence documents [18]. By pretraining on corpora of conflict news and reports, ConfliBERT acquires domain-specific understanding that enables it to outperform generic language models (like standard BERT) on tasks such as event classification and protest detection [18], [19]. This allows to use ConfliBERT to automatically extract structured event data (e.g. who did what to whom) from unstructured text with higher accuracy, and improving the timeliness and quality of event feeds that other warning models can rely on.

Another cutting-edge approach is leveraging network and deep learning methods to forecast conflict dynamics among specific actors. A recent study by Croicu and von der Maase combines news text with structured event data to predict escalations and de-escalations in violence between pairs of actors [20]. This project uses transformer neural networks to create vector representations of actors based on their context in news articles, and then uses those embeddings to anticipate conflict dyad dynamics (e.g. government–rebel interactions). The result is a model that can detect volatile shifts in conflict at a granular level – essentially mapping how relationships between warring parties evolve in real time. Initial results show this method yields more prompt and actor-specific warnings of escalation and deescalation dynamics, surpassing the predictive power of traditional structural country-level models in capturing short-term changes in a conflict dynamics.

Additionally, various operational early-warning systems developed by governments and NGOs. A recent comparative review by Hegre et al. outlines a landscape of at ten prominent systems [6]. Some are academic or nonprofit projects akin to ViEWS and Conflict Forecast, while others are run by policy organizations and are geared toward direct conflict prevention on the ground.

For example, PREVIEW is an closed-source early-warning program of the German Federal Foreign Office that combines quantitative risk modeling with qualitative expert analysis. PREVIEW analysts produce country risk forecasts using statistical models with human in the loop augmentation by regional desks' assessments.

The Integrated Crisis Early Warning System (ICEWS), originally developed by DARPA and now maintained by a U.S. defense contractor, is a another example of a closed-source system that continuously monitors global event data to predict crises [6], [21]. ICEWS uses a hybrid method that combines automated event coding, sentiment analysis, and ensemble forecasting. It covers hundreds of countries, but its inner workings and outputs are confidential, intended for intelligence and military use.

Moveover, there are several impressive NGO-led and academic initiatives: the ACLED Volatility Risk Index (VRI), for instance, provides short-term forecasts of conflict surges by analyzing deviations from baseline violence levels in ACLED's real-time event data [22]. The Early Warning Project (EWP), run by the U.S. Holocaust Memorial Museum, takes yet another angle – it produces annual risk assessments specifically for mass atrocities, ranking countries by their estimated probability of experiencing a new mass-killing episode [23]. The EWP model is a small statistical model (a regularized logistic regression on around 20 predictors) that is updated yearly and made to encourage preventive action in the highest-risk countries.

Other systems like Peoples Under Threat by Minority Rights Group International and the Atrocity Forecasting Project by Australian National University also publish risk indices for violence against civilians or mass atrocities, typically on an annual or semi-annual cycle [6].

Each system has its own methodological nuances. For instance, the EU's Global Conflict Risk Index uses a composite indicator approach with 24 variables to score country risk [24], whereas Water, Peace & Security integrates environmental and water stress data into machine-learning models for conflict in developing regions. A key distinction remains between academic/open systems and operational/closed ones: academic projects generally publish their forecasts and methods, while operational systems (like ICEWS or UN/AU regional models) keep their analytics and findings internal [6].

Finally, despite the clear societal benefits that accurate early conflict warning systems can provide, the literature increasingly highlights significant reproducibility and collaboration challenges in this field. Many projects label themselves as transparent or open-source, yet in practice it often remains difficult to fully replicate their results or effectively integrate insights across different modeling approaches [6], [25]. Even within academia, datasets may be partially restricted and most important parts of the code may not be shared. For example, although initiatives such as ViEWS and Conflict Forecast have made significant progress, partial openness and limitations on data accessibility hinder the broader scientific community's efforts to verify, reproduce, and build upon these systems. This lack of openness and reproducibility creates a substantial obstacle for the fast advancements in the early conflict forecasting domain.

In summary, despite significant advancements, existing early conflict warning systems still suffer from substantial reproducibility and transparency challenges, primarily due to partial openness and restricted data availability. Moreover, traditional structural models, while useful, often fail to effectively capture short-term triggers and nuanced socio-political dynamics of emerging conflicts. This thesis explicitly addresses these identified gaps by developing a fully open-source conflict forecasting framework that combines publicly accessible datasets with advanced machine learning methods. Through systematic evaluation of multiple state-of-the-art modeling techniques and transparent provision of all developed resources, this research aims to set a new standard for reproducibility and model performance in the early conflict forecasting domain.

3

Datasets

Comprehensive political datasets are challenging to assemble and are very limited in supply. This work uses two publicly available datasets published by ViEWS and Conflict Forecast organizations. Both of these datasets contain a diverse set of features and are used in the Early Conflict Forecasting domain. Additionally, we create our own dataset, which is a full reproduction of Conflict Forecast Historical dataset. We do this to extend the amount of data available for training the models, as Conflict Forecast only shares the reduced dataset since 2010. Table 3.1 compares all three datasets we use in this work.

All datasets used in this study are on a country-month level and source the fatalities features from the Uppsala Conflict Data Program (UCDP) dataset.

	ViEWS	**Conflict Forecast**	**UCDP-based Reproduction**
Data Since	1990-01	2010-01	1989-01
Data Till	2024-01	2024-12	2024-12
News Features	No	Yes (15 dimensions)	No
# Features	126	41	26
# Rows	77,300	30,500	71,918

Table 3.1: Overview of Datasets used in this work: ViEWS, Conflict Forecast, and our custom reproduction UCDP-based of Conflict Forecast Historical dataset

3.1. ViEWS dataset

The ViEWS dataset is composed of multiple sources with structural variables such as indicators about economics, policy, and country terrain. The dataset covers 213 unique country IDs from Jan 1990 to Jan 2024 and contains 77,300 rows. The dataset composes features from the World Development Indicators [26] and V-Dem [27], data on politically excluded ethnic groups [28], demographic factors [29], protests from ACLED [30], and national water resources from AQUASTAT [31]. Additionally, the ViEWS team engineers temporal and spatial lag features with decays to enhance the predictive power of the dataset. The total number of features in the dataset is 126.

3.2. Conflict Forecast dataset

Conflict Forecast builds their dataset based on fatality counts from UCDP, population data from the World Bank and also adds news topic modelling features based on LDA modeling of the Factiva news database [16]. The publicly released reduced dataset covers 170 countries from January 2010 through April 2024, comprising approximately 32 000 rows and 41 features per row. It includes fifteen monthly topic-share variables, rolling and discounted past-fatality indicators, months-since-conflict counters, and static covariates like population and ongoing-conflict flags [16].

3.3. Pipeline for UCDP dataset

The UCDP dataset is our custom dataset built based on the fatality data from the UCDP Georeferenced Events Dataset and by closely following the approach taken by the Conflict Forecast team. Effectively, this dataset reproduces the historical Conflict Forecast dataset without the news features. The dataset covers 35 years of data from Jan 1989 to April 2024 for 170 unique country IDs and contains 72,485 rows.

To achieve this, we build the 16-stage pipeline to fetch the latest UCDP data, aggregate it to the country-month level and calculate the features following the Conflict Forecast definitions. We calculate similarity score over the overlapping data with the Conflict Forecast dataset to ensure that the data is calculated in the same way. The similarity score for each feature group is shown in Table 3.2 and the overall similarity score calculated over all features with tolerance of $\pm 10^{-5}$ is 71.41%.

Feature Group	**Similarity (%)**
past bestpc	78.35
since	100
neighbors	90.67
discounted	96.51
ongoing	100
population	100
target group (ons_armedconf_12_target)	100
fatalities_UCDP	100
armedconf	100

Table 3.2: Similarity percentage for each feature group between the UCDP dataset and the Conflict Forecast dataset. Scores are based on overlapping data from January 2010 to December 2023, with a tolerance of $\pm 10^{-5}$.

The description of each pipeline step is provided in the Appendix A. The implementation of the pipeline is available in the `create_ucdp_dataset` folder of the repository.

4

Methodology

This section describes the datasets used in this work, the definition of the dependent variable, the models used, and the training approach of our machine learning models.

4.1. Definition of the dependent variable

The dependent variable for our Early Conflict Warning forecasting task is constructed in a way the model would output the maximum risk of the armed conflict in the next 12 months. This makes the task a binary classification task, where any armed conflict in the next 12 months is considered a positive instance and no armed conflict is considered a negative instance.

We follow the Conflict Forecast approach to define the armed conflict occurrence and the dependent variable. An armed conflict is said to occur in the country c at time t if the number of battle-related fatalities per 1 million inhabitants exceeds a threshold of 0.5 in a given month.

$$\text{conflict}_{c,t} = \begin{cases} 1 & \text{if } \frac{\text{fatalities}_{c,t}}{\text{population}_{c,t}} \cdot 10^6 > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

To construct the dependent variable $y_{c,t}$ for model training that can predict conflicts for 12 months ahead, we annotate the time series as follows:

- $y_{c,t} = 1$ if a conflict onset occurs in the next 12 months after time, t and it was preceded by at least one full month of peace.
- $y_{c,t} = 0$ if there is no conflict in the next 12 months, and the country is not in an ongoing conflict.
- $y_{c,t} = \text{NaN}$ if the country is already in an ongoing conflict at time t.

The exception for the NaN values was introduced by Conflict Forecast to make the model focus more on forecasting the conflict onsets rather than learning the patterns of ongoing conflicts. We do not use such instances for training and evaluation of the model.

We use this definition of the armed conflict-dependent variable throughout the entire project. Importantly, this formulation implies that the dependent variable $y_{c,t}$ cannot be computed for the last 12 months of the data available, since for each time point t, it requires knowledge of conflict events in the future 12-month window $(t+1, \ldots, t+12)$. This will be important to avoid leaking data about the future when evaluating model performance described in the section 4.4, where we describe the model training approach.

4.2. Definition of hard onsets

In addition to evaluating all conflict onsets, we pay particular attention to so-called *hard onsets*, a subset of conflict outbreaks that are especially challenging to forecast. Following the definition used by the Conflict Forecast team, a hard onset refers to the start of a new armed conflict in a country that has experienced at least 60 consecutive months (i.e., 5 years) of peace. These cases are of high policy relevance, as they typically represent unexpected escalations in historically stable countries—situations where timely early warnings can have the greatest preventative impact.

By isolating these rare but critical events, we aim to assess the ability of models not only to detect general conflict patterns but to anticipate the emergence of new cycles of violence in seemingly peaceful contexts.

4.3. Models

This section introduces the machine learning models employed in our analysis and outlines the approaches used to train them. Specifically, we explore three state-of-the-art machine learning methods: XGBoost, the AutoGluon AutoML framework, and the recently introduced transformer-based TabPFN model. Each of these models has demonstrated strong performance across various predictive tasks and represents different approaches to machine learning—gradient boosting, automated model selection and ensembling, and transformer-based in-context learning.

Due to the highly imbalanced nature of our datasets, as highlighted in Section 3, traditional accuracy metrics are insufficient and overly optimistic, given the predominance of non-conflict instances. Therefore, following established practice in conflict forecasting literature [11], [15], we use Precision-Recall Area Under the Curve (PR-AUC) as our primary performance metric. All models are trained and optimized to maximize PR-AUC.

4.3.1. XGBoost

Extreme Gradient Boosting (XGBoost) is a highly efficient and scalable implementation of the gradient boosting framework, which has demonstrated state-of-the-art performance across a wide range of machine learning tasks [32]. The algorithm operates by sequentially building an ensemble of weak learners, typically decision trees, where each new tree is trained to correct the residual errors of the preceding models. By optimizing a regularized objective function that combines a loss function and a penalty for model complexity, XGBoost effectively mitigates overfitting and enhances generalization. Its computational efficiency is achieved through techniques such as a specialized tree-finding algorithm and parallel processing capabilities. Due to its robustness and predictive power, XGBoost has been widely adopted in academic research and industry, including in the domain of conflict forecasting [12].

Hyperparameter	Range	Description
`learning_rate`	[1e-5, 0.1]	Step size shrinkage to prevent overfitting.
`n_estimators`	[50, 500]	Total number of boosting rounds.
`subsample`	[0.1, 1.0]	Fraction of training data instances sampled per tree.
`colsample_bytree`	[0.1, 1.0]	Fraction of features sampled when constructing each tree.
`colsample_bylevel`	[0.1, 1.0]	Fraction of features sampled for each tree level.
`colsample_bynode`	[0.1, 1.0]	Fraction of features sampled for each node split.
`reg_alpha`	[0.01, 100]	L1 regularization term on model weights.
`reg_lambda`	[0.01, 100]	L2 regularization term on model weights.
`gamma`	[0.01, 100]	Minimum loss reduction required to make a partition.
`max_depth`	[3, 10]	Maximum depth of an individual tree.
`min_child_weight`	[0.01, 10]	Minimum sum of instance weight needed in a child node.

Table 4.1: XGBoost Hyperparameters for Optimization

In the context of early conflict warning systems, XGBoost is used for modeling the complex, non-linear relationships between structural predictors and the likelihood of political violence [24]. Systems such as the ViEWS employ XGBoost, often within a broader ensemble of machine learning models, to generate forecasts of conflict-related fatalities [12]. To maximize the predictive accuracy of the XGBoost model, its hyperparameters must be carefully calibrated. Table 4.1 lists the hyperparameters we optimize and their ranges for XGBoost model. We use Optuna, a modern hyperparameter optimization framework that employs Bayesian optimization techniques to efficiently search large parameter spaces and identify the optimal model configuration [33].

We train the XGBoost model on all three datasets: ViEWS, Conflict Forecast, and UCDP with a computational budget of 5 hours per dataset per year.

4.3.2. AutoGluon

AutoGluon is an open-source Automated Machine Learning (AutoML) framework designed to achieve state-of-the-art performance across various machine learning benchmarks [34]. It streamlines the machine learning pipeline by automatically managing tasks such as model selection, feature engineering, and hyperparameter tuning. One of AutoGluon's key strengths is its use of multi-layer stack ensembling. In this approach, AutoGluon simultaneously trains multiple base models, whose predictions are then combined into a final, weighted ensemble model. This method leverages the complementary strengths of diverse algorithms, often resulting in a more robust and accurate predictive model compared to any single manually optimized model.

For the early conflict forecasting task, we specifically utilize the AutoGluon-Tabular module, which is tailored for structured data. This module automatically evaluates a comprehensive range of machine learning models, including tree-based ensembles and neural networks. Table 4.2 summarizes the models supported by AutoGluon-Tabular that we use through the study.

Model	**Description**
`LightGBM (LGBM)`	A high-performance gradient boosting framework using histogram-based algorithms for speed and efficiency.
`CatBoost`	A gradient boosting algorithm that excels at natively handling categorical features without extensive preprocessing.
`XGBoost`	A scalable and regularized gradient boosting implementation known for state-of-the-art performance.
`Random Forest (RF)`	An ensemble method building multiple decision trees on random subsets of data and features to improve accuracy.
`Extra Trees (XT)`	A variant of Random Forests that introduces more randomness in how node splits are chosen to reduce variance.
`TabularNeuralNetTorch`	A PyTorch-based multi-layer perceptron (MLP) specifically designed for tabular data.
`NNFastAiTabular`	A neural network for tabular data built using the fast.ai library, featuring useful defaults.

Table 4.2: Models Supported by AutoGluon-Tabular

The main feature of AutoGluon is its multi-layer stacking and ensembling process, which allows to maximize the predictive performance of the models by using so called *wisdom of the crowd* [11]. The multi-layer stacking process begins by training base models in parallel on the initial data layer. These initial models generate out-of-fold predictions, which serve as meta-features for subsequent model layers. At each subsequent layer, AutoGluon trains additional models and fits a weighted ensemble, identifying the optimal linear combination of model predictions. Consequently, the final model is typically a comprehensive weighted ensemble derived from all successful models across multiple layers.

We train AutoGluon models for all three datasets and allocate an identical computational budget of 5 hours per dataset per year to ensure fairness and consistency in comparing AutoGluon's performance with other models.

4.3.3. AutoGluon Hyperparameter Optimization

By default, AutoGluon does not perform hyperparameter optimization (HPO) for its constituent models. Instead, it selects hyperparameters from a predefined portfolio of model configurations and trains them within the provided computational budget. While this portfolio-based approach typically provides robust performance across various tasks, it may not be optimal for highly specialized or domain-specific datasets such as the one we have for early conflict forecasting.

To further improve the results, we explored AutoGluon's built-in hyperparameter optimization feature, which optimizes the model hyperparameters instead of relying solely on predefined portfolio. We initially considered all models supported by AutoGluon-Tabular listed in Table 4.2, but through our experimentation, we identified that hyperparameter optimization brought substantial improvement only for Random Forest and CatBoost models. Other models, such as XGBoost, LightGBM, and various Neural Networks, did not demonstrate significant performance gains with enabled HPO during our experimentation.

Due to computational constraints, we conducted hyperparameter optimization exclusively on the UCDP dataset, which was identified as the most promising for achieving state-of-the-art performance. The optimization was integrated into the yearly training scheme described in Section 4.4. This means that for every year we make a separate search for the most optimal hyperparameters. The approach for HPO does not deviate from our yearly training scheme definition and has precisely the same training, validation, and test dataset splits.

Table 4.3 summarizes the hyperparameter ranges we set for the Random Forest and CatBoost models during HPO, and the best value of of the range is selected for each prediction year. We employed Bayesian optimization as our hyperparameter search strategy, setting a computational budget of 30 minutes per model per year. The performance metric to optimize was set to Precision-Recall AUC.

Model	**Hyperparameter**	**Search Space**
Random Forest	`n_estimators`	[300, 400]
	`max_depth`	[3, 5]
CatBoost	`scale_pos_weight`	[5.0, 6.0]
	`iterations`	[500, 2000]
	`depth`	[6, 10]
	`learning_rate`	[0.005, 0.1] (log-scale)
	`l2_leaf_reg`	[1, 10]
	`border_count`	[32, 255]
	`random_strength`	[1, 20]
	`bagging_temperature`	[0, 1]
	`early_stopping_rounds`	20 (fixed)

Table 4.3: Hyperparameter search spaces and tuning settings for Random Forest and CatBoost models. Each model was tuned for 30 min per year, optimizing PR-AUC.

4.3.4. TabPFN

TabPFN v2 is a Transformer-based tabular foundation model designed for in-context learning on small and medium-sized tabular datasets, enabling classification and regression without explicit model training or hyperparameter tuning for each new task [35]. Its architecture allows the model to process the entire dataset as a prompt, performing accurate predictions in a single forward pass. TabPFN's training dataset consists exclusively of synthetic tables generated by structural causal models. This is great for conflict forecasting domain, as this ensures that no real-world data is leaked into the training process, to give an unfair advantage to the model. Empirical evaluations demonstrate that TabPFN significantly outperforms traditional gradient-boosted methods like CatBoost, XGBoost and even AutoGluon framework, achieving superior predictive performance with drastically reduced inference latency. The main limitation of TabPFN is that it was trained on the datasets below 10,000 rows, and authors explicitly state that the performance is not guaranteed on larger datasets.

Due to the novelty of this model, there was no prior work on applying it to Early Conflict Forecasting. In this work, we apply TabPFN to the Conflict Forecast dataset only as it is the smallest of all available. We also briefly test the performance on two larger ViEWS and UCDP datasets. Notably, despite being the smallest out of all three datasets, the Conflict Forecast dataset size is at least three times larger than the maximum recommended size for TabPFN [36]. As per computational budget, since TabPFN does not require model training, we do not perform any hyperparameter optimization and only run the model without upper training limit.

To improve the TabPFN performance even further, the AutoTabPFN was introduced [37]. AutoTabPFN implements a post-hoc ensemble strategy by instantiating multiple perturbed TabPFN base models within a user-specified time budget and greedily combining their outputs into a weighted ensemble optimized for a chosen metric. Empirical results on over 300 datasets show that even modest ensemble sizes yield consistent, statistically significant accuracy improvements [37].

Our initial assumption was that even a short computational budget for AutoTabPFN ensemble would surpass the vanilla TabPFN model. In the Section 6.1, we found that the 5 and 15-minutes AutoTabPFN ensemble performed worse than the vanilla TabPFN model on the ViEWS dataset. As AutoTabPFN is a computationally expensive process, this spiked our interest in further investigating the optimal computational budget for the AutoTabPFN ensemble for our subsequent datasets.

We conducted a computation budget test on the Conflict Forecast dataset, varying the runtime from 5 minutes up to one hour. The results are shown in Figure 4.1. We observed that the performance gradually increases and peaks at 45 minutes before plateauing and even dipping slightly at longer runtimes. Interestingly, the performance of AutoTabPFN ensemble is weaker than vanilla TabPFN before the 21-minute mark for Conflict Forecsat and also slightly decreases at 1 hour mark. This finding mirrored the evidence from Section 6.1 and based on these results, we set 45 minutes as the computational budget for all subsequent AutoTabPFN experiments.

4.4. Model training approach

This subsection outlines the training strategy adopted across all experiments, including key validation design choices, measures taken to prevent data leakage and further optimizations used to improve model accuracy. We follow the standard pseudo-out-of-sample evaluation approach widely used in the conflict forecasting literature. Under this setup, models are trained exclusively on data that would have been available at the time of prediction, thereby simulating realistic forecasting conditions.

As discussed in Section 4.1, our dependent variable $y_{c,t}$ is defined based on conflict occurrence in

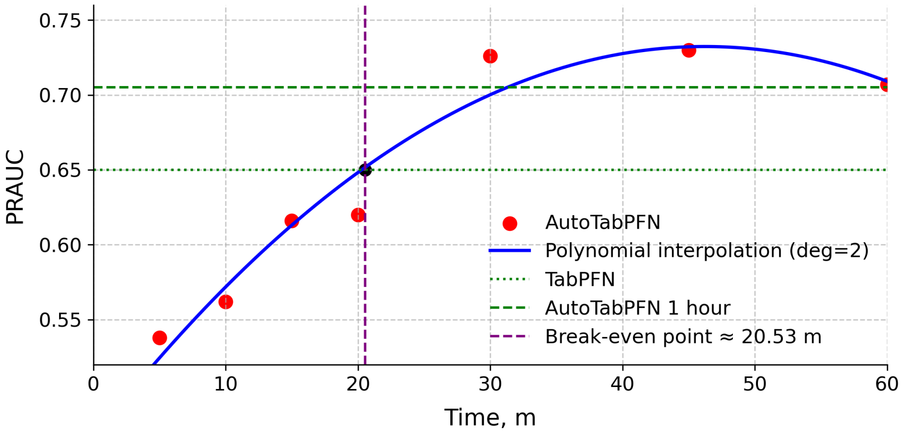


Figure 4.1: Performance of AutoTabPFN ensemble on Conflict Forecast dataset with different computational budgets with Jan 2022 as a test set.

the 12-month future window, $(t+1,\ldots,t+12)$. Consequently, we cannot compute our dependent variable for the last available 12 months of each dataset, since those future observations are unavailable. For example, if were currently in January 2024 and wished to train a model to forecast the risk of armed conflict over the next 12 months, we would have had observations up until January 2024 with the dependent variable only up to January 2023, as we do not have data for February 2024 to construct the dependent variable for February 2023. Thus for pseudo-out-of-sample evaluations, in case we want to simulate that we are now in January 2024, we must limit the training data to observations no later than January 2023. Including data beyond that point would imply having knowledge of conflict events that happened beyond January 2024, which violates the causal structure of forecasting and would leak the information about the future into the model.

In this work, we apply two model training schemes: yearly and monthly. They differ in how the training data is constructed and how frequently models are updated. These approaches, along with validation set design and additional optimization procedures, are described in detail in the following subsections.

4.4.1. Yearly training scheme

Under the yearly training scheme, we retrain a new model for each forecast year, using only data that would have been available prior to that year. This setup ensures that the evaluation is pseudo-out-of-sample and prevents any information from leaking from the future into the model. Crucially, to ensure no data leakage into the future, we introduce an 11-month buffer between the end of the validation set and the beginning of the test period. The visualization of the yearly training scheme is shown in Figure 4.2.

Each yearly model is trained on a dedicated training set and validated on a temporally separated validation set. Once the optimal hyperparameters are selected based on validation performance, the model is retrained on the combined training and validation data.

This process is repeated independently for each year as we do not reuse model weights or

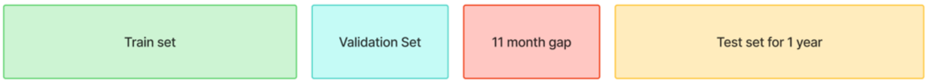


Figure 4.2: Yearly training scheme for the Early Conflict Forecasting task. Each year, a new model is trained on data available up to that point, with an 11-month buffer to prevent leakage of data about future state of the dependent variable.

hyperparameters across forecast years. A separate model is trained for each forecast window (e.g., 2018, 2019, . . . , 2025), allowing the training procedure to adapt fully to the data available up to that point. This differs from the methodology used by the Conflict Forecast team, who train a single model once and reuse it across years, retaining fixed hyperparameters. Pseudocode for the yearly training scheme is outlined in Algorithm 1.

Algorithm 1 Yearly Forecast: Pseudo Out-of-Sample Training.

```
Require: Full monthly sample $D = \{d_{1990\text{m}1}, d_{1990\text{m}2}, \dots, d_{2025\text{m}04}\}$
Require: Set of forecast years $\mathcal{Y} = \{2010, \dots, 2024\}$   ▷ dependent variable is availalbe only till 2024-04
Require: Validation window length $V$ (years)
Require: Forecasting model class $F$
Ensure: Predictions $\hat{y}_{\text{Jan–Dec } Y}$ for all $Y \in \mathcal{Y}$
for each $Y \in \mathcal{Y}$ do
    $test\_input \leftarrow \{d_{(Y-1)\text{m}1}, \dots, d_{(Y-1)\text{m}12}\}$
    $holdout \leftarrow \{d_{(Y-2)\text{m}2}, \dots, d_{(Y-2)\text{m}12}\}$   ▷ 11-month gap
    $val\_end \leftarrow d_{(Y-2)\text{m}1}$   ▷ Jan $(Y-2)$
    $val\_start \leftarrow val\_end - (12V)$ months   ▷ Jan $(Y-2-12V)$
    $TRAIN \leftarrow \{d_t \mid 1990\text{m}1 \le t < val\_start\}$
    $VALID \leftarrow \{d_t \mid val\_start \le t \le val\_end\}$
    $\theta^\star \leftarrow \arg\min_{\theta} \text{Loss}\big(F(\theta, \text{TRAIN}), \text{VALID}\big)$   ▷ hyper-parameter optimisation
    Refit model $F$ on TRAIN ∪ VALID with $\theta^\star$
    $\hat{y}_{\text{Jan–Dec } Y} \leftarrow F(test\_input)$
end for
return $\{\hat{y}_{\text{Jan–Dec } Y}\}_{Y \in \mathcal{Y}}$
```

4.4.2. Monthly training scheme

To mimic the information set available to an operational warning system, we also implement a rolling, month-by-month training procedure inspired by the one implemented by Conflict Forecast. At the beginning of each calendar month, the one more month of data with the dependent variable becomes available, and we augment the training set with this new data point. Opposed to the yearly training scheme, where the model is trained once per year, the monthly training scheme allows us to readjust the model every month, which allows the model to adapt to the most recent changes in the data. Such a monthly model rebuilding and recalibration is a common practice in the conflict forecasting literature [15], [38].

The structure of the data split, illustrated in Figure 4.3, is as follows. The training window begins in January 1990 and expands by exactly one month at every iteration; immediately after the training set follow a validation set windows, which we keep of fixed length. The validation block always ends twelve months before the forecast origin, so an eleven-month hold-out buffer separates the last validated instance from the target period.

For computational efficiency, hyperparameters are chosen once, in the January run of each calendar year, by minimising the loss on that year's validation block. They are then kept fixed for the remaining eleven monthly refits, so that the only change from one iteration to the next is the inclusion of the newly available training month and retraining of the model on the newly available data.

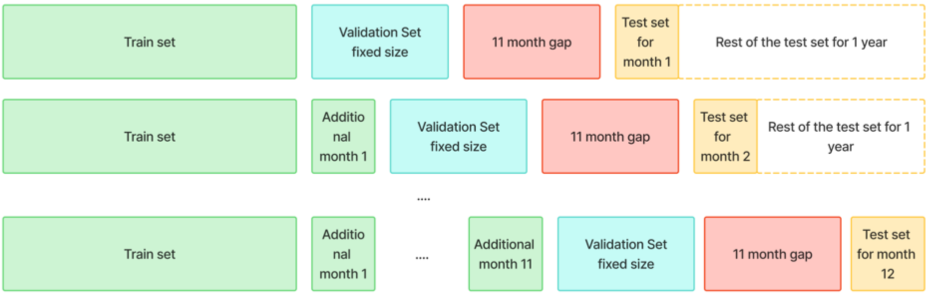


Figure 4.3: Monthly training scheme for the Early Conflict Forecasting task. Each month, a new model is retrained on data available up to that point, with fixed hyperparameters chosen in the first month of each year. An 11-month buffer is maintained to prevent leakage of data about the future state of the dependent variable.

Compared with the yearly retraining schedule, the monthly scheme therefore produces twelve independent refits per calendar year, each based on a slightly larger estimation sample. In doing so, it allows the learner to absorb the most recent conflict signals. While the assumption is that monthly training scheme will benefit from having additional data, in the Section 6 we find that the improvement is marginal and sometimes even negative.

Algorithm 2 Monthly Forecast: Rolling Refit with Fixed Hyper-parameters.

```
Require: Full sample $D = \{d_{1990\mathrm{m1}}, \dots, d_{2025\mathrm{m04}}\}$
Require: Forecast years $\mathcal{Y} = \{2010, \dots, 2024\}$
Require: Validation length $V$ (years)
Require: Forecasting model class $F$
Ensure: Predictions $\hat{y}_{Y,m}$ for every $Y \in \mathcal{Y}$ and $m \in \{1, \dots, 12\}$
for each $Y \in \mathcal{Y}$ do                                        ▷ loop over calendar years
    $val\_end \leftarrow d_{(Y-2)\mathrm{m1}}$                                         ▷ Jan $(Y-2)$
    $val\_start \leftarrow val\_end - (12V)$ months                            ▷ $V$ full years
    $TRAIN \leftarrow \{d_t \mid 1990\mathrm{m1} \le t < val\_start\}$
    $VALID \leftarrow \{d_t \mid val\_start \le t \le val\_end\}$
    $\theta^\star \leftarrow \arg\min_\theta \mathrm{Loss}\big(F(\theta, \mathrm{TRAIN}), \mathrm{VALID}\big)$      ▷ hyper-parameter optimisation
    for $m = 1$ to 12 do
        $test\_input \leftarrow \{d_{(Y-1)\mathrm{m}m}\}$                           ▷ current forecast origin
        Refit $F$ on TRAIN $\cup$ VALID with fixed $\theta^\star$
        $\hat{y}_{Y,m} \leftarrow F(test\_input)$
        if $m < 12$ then                                     ▷ prepare windows for next month
            $val\_start \leftarrow val\_start + 1$ month
            $val\_end \leftarrow val\_end + 1$ month
            $TRAIN \leftarrow \{d_t \mid 1990\mathrm{m1} \le t < val\_start\}$       ▷ TRAIN grows by one month
            $VALID \leftarrow \{d_t \mid val\_start \le t \le val\_end\}$ ▷ VALID slides by one month and stays of fixed length
        end if
    end for
end for
return $\{\hat{y}_{Y,m}\}$
```

4.4.3. Optimal validation set size

Selecting an appropriate validation window length is critical to balance bias and variance in our pseudo-out-of-sample evaluation. Following the hold-out cross-validation approach described by Lynam et al. [25], we experimented with multiple amount of years in the validation set (see Algorithm 1) to identify the configuration that maximizes Precision–Recall AUC on the validation and test sets.

To conduct our experiment, we run train a yearly AutoGluon model for 1 hour on the best configuration for each combination of validation window length V and dataset. Due to computational limitations, the ViEWS and Conflict Forecast validation window lengths were chosen mostly empirically, while the UCDP validation window length was chosen based on a systematic evaluation of multiple validation window lengths.

We stay consistent with our previous work, and we put 4 years in the validation set for ViEWS dataset. We find that due to the high dimensionality in the ViEWS dataset, 4 years of validation set allows to assign much more data to the training set, which allows the model to learn better.

To select the optimal validation window length for the Conflict Forecast, we conducted an empirical evaluation of multiple validation window lengths V ranging from 1 to 6 years. We find that 4 validation years on Conflict Forecast V allow for the best balance between the amount of the data in the validation set and the amount of the data being left for training.

The UCDP dataset has much fewer dimensions than ViEWS and has 20 years more data than Conflict Forecast. This allows using a longer validation set, to stabilize the model performance, while also keeping the complexity of the model low because of fewer dimensions compared to ViEWS dataset. For the UCDP dataset, due to computational limitations, we empirically test the validation window lengths from 1 to 14 years. We find that the higher validation set size improves the out-of-sample performance, so then we systematically test the validation set sizes for $V \in \{9, 12\} \cup \{14\}$ years. The summary of the results is shown in Table 4.4. We choose the 10-year validation window as it achieves the highest mean test PR AUC set with reasonable stability.

Validation Years (V)	Mean Test PR AUC	Std Dev	Corr (Val vs Test)
9	0.7074	0.0775	–0.116
10	**0.7181**	**0.0692**	**–0.402**
11	0.7028	0.0720	–0.511
12	0.7176	0.0582	–0.260
14	0.7021	0.0548	–0.362

Table 4.4: UCDP: Validation window analysis. The 10-year window (bolded) achieves the highest mean test PR AUC with reasonable stability.

Table 4.5 report the chosen validation set size for each dataset. All subsequent experiments adopt these optimal values of V for their respective datasets.

Dataset	Amount of validation years
ViEWS	4
Conflict Forecast	4
UCDP	10

Table 4.5: Optimal validation set size for each dataset, chosen to maximize Precision–Recall AUC.

5

Modeling Enhancements

Ensemble methods are a powerful way to boost predictive performance by combining multiple models. This section outlines four key improvements we tested to enhance our conflict forecasting performance: stacking, ensembling, combining stacking and ensembling, and news-feature experiments.

5.1. Stacking

Model stacking, or stacked generalization, is a meta-learning strategy designed to leverage the complementary strengths of multiple predictive models. Rather than settling on a single algorithm, stacking unfolds in two distinct steps. In the first tier, a collection of base learners is trained on the original feature set; each base learner generates its own out-of-sample predictions via a cross-validation or rolling-forecast scheme, thereby guarding against information leakage. These first-stage predictions encapsulate diverse patterns of the data that individual models have learned. In the second tier, a meta-learner is then trained on this expanded feature space, which consists of the original covariates (optionally) plus the level-one predictions. By doing so, the meta-learner learns how to optimally weight and combine the strengths of each base model, often yielding improvements in calibration, discrimination, and overall predictive accuracy relative to any single constituent.

In our application to monthly armed conflict risk prediction, we adopt precisely this two-stage architecture and Figure 5.1 illustrates our stacking pipeline in detail. At Level 1, an AutoGluon RF&CAT (Random Forest and Cat Boost) model is trained each year on UCDP-derived structural covariates spanning 1990–2023. We then generate truly out-of-sample predictions for every country–month in 2010–2023 via a year-by-year rolling scheme, ensuring that each forecast is based only on data that would have been available at the time. The baseline armed conflict risk score, which synthesize long-run historical signal, is then merged with the publicly accessible Conflict Forecast dataset's with LDA news and structural features (our UCDP dataset and Conflict Forecast dataset are almost identical for the 2010–2023 period, so the results do not change on whether we merge Conflict Forecast news features to UCDP dataset or our Model 1 prediction to Conflict Forecast dataset). At Level 2, a second AutoGluon RF&CAT model learns to map this enriched feature vector—combining structural history and real-time news metrics—onto coup-risk outcomes for 2018–2023. Empirically, this stacked

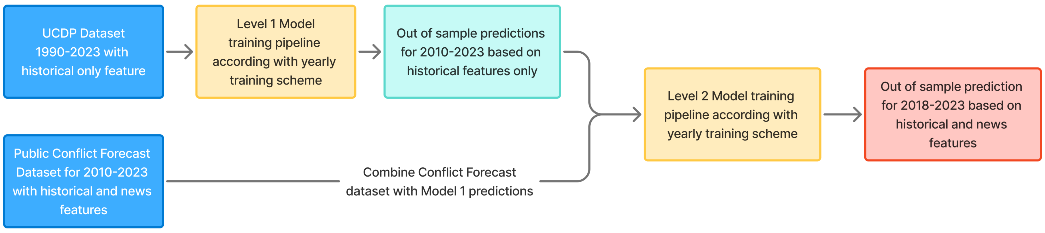

UCDP Dataset 1990-2023 with historical only feature
Level 1 Model training pipeline according with yearly training scheme
Out of sample predictions for 2010-2023 based on historical features only
Public Conflict Forecast Dataset for 2010-2023 with historical and news features
Combine Conflict Forecast dataset with Model 1 predictions
Level 2 Model training pipeline according with yearly training scheme
Out of sample prediction for 2018-2023 based on historical and news features

Figure 5.1: Our two-stage stacking pipeline. Level 1 distills historical conflict features into an out-of-sample baseline forecast of armed conflict; Level 2 refines that forecast by incorporating Conflict Forecast news indicators that are available only since 2010.

ensemble improves the precision recall score for all onsets, confirming that the addition of the news indicators can add another 1–2 percent to the overall performance of the model.

5.2. Ensembling

Ensembling is an effective modeling approach that combines predictions from several individual models to create a single, more robust forecast. This method is widely used in the field of Early Conflict Forecasting, and often referred as a *wisdom of the crowd* approach [11]. Typically, each model's performance is first evaluated on a validation set, and then based on additional data between the validation and test sets the model predictions are used to derive and calibrate weights to reflect how much each model should contribute to the final ensemble. Thus, predictions from better-performing models have greater influence, while weaker models contribute less.

In this work, we use ensembling extensively. First, we utilize AutoGluon's built-in weighted ensemble capabilities, where the framework automatically combines the predictions of its internal models based on validation performance (see Section 4.3.2). Additionally, we implement a custom weighted ensembling pipeline to combine multiple AutoGluon final models. This approach allows us to leverage the strengths of various individually trained AutoGluon models, further boosting predictive performance. The exact implementation and performance gains from this custom ensemble approach are discussed in detail in the Section 5.3. Figure 5.2 visualizes the concept of weighted ensembling.

5.3. Combining Stacking and Ensembling

To further enhance predictive performance, we combine stacking and ensembling into a unified modeling pipeline. Figure 5.3 illustrates this combined approach.

The stacking part of our pipeline follows the definition outlined in Section 5.1, where we first train a Level-1 AutoGluon Random Forest and CatBoost model on historical-only features from the UCDP dataset, and subsequently enrich these predictions with news-based features

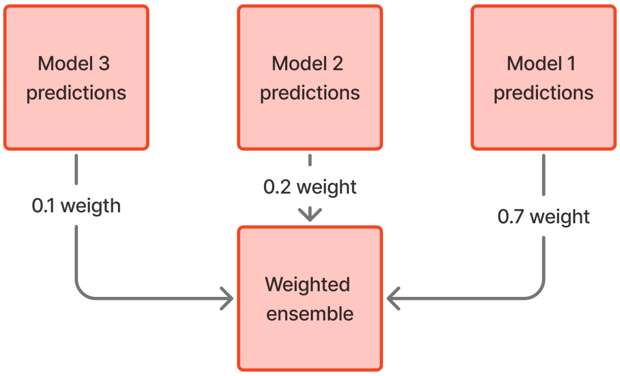


Figure 5.2: Weighted ensembling approach: each model's predictions are combined into a single forecast using validation-derived weights.

provided by the Conflict Forecast dataset. We find this pipeline performs strongly on all conflict onsets, but is less robust when predicting hard onsets.

To improve the stacked model further, we employ a weighted ensembling technique by integrating an additional AutoGluon Random Forest model trained solely on historical features from the UCDP dataset. This AutoGluon HPO model performs strong on hard onsets, which makes it a great candidate for ensembling with the stacked model (see Section 4.3.3 for the model definition). By combining these two complementary models through weighted ensembling, we leverage both the historically and news informed stacked model and the simpler, yet robust, historical-only model. We conduct a systematic search to identify optimal ensemble weights based on out-of-sample performance on validation set. The optimal weights derived for the final ensemble model are 0.25 for the stacked AutoGluon model with news-enhanced predictions and 0.75 for the historical-only AutoGluon Random Forest model.

This combination yields our strongest performing model, achieving our highest accuracy on all conflict onsets and a decent performance on hard onsets (see Section 6.3 for detailed performance analysis). It is important to note that while this combined approach maximizes predictive accuracy, it introduces additional complexity in terms of explainability. Unlike simpler individual models such as XGBoost or single AutoGluon models that have built-in support for SHAP value analysis, this ensemble pipeline requires additional steps to derive SHAP values. Nevertheless, if interpretability is required, SHAP analysis can still be feasibly conducted for this pipeline.

5.4. News-feature experiments

We conduct experiments with text features to enhance our models with additional signals from monthly news summaries. Unlike the Conflict Forecast approach, which relies on LDA

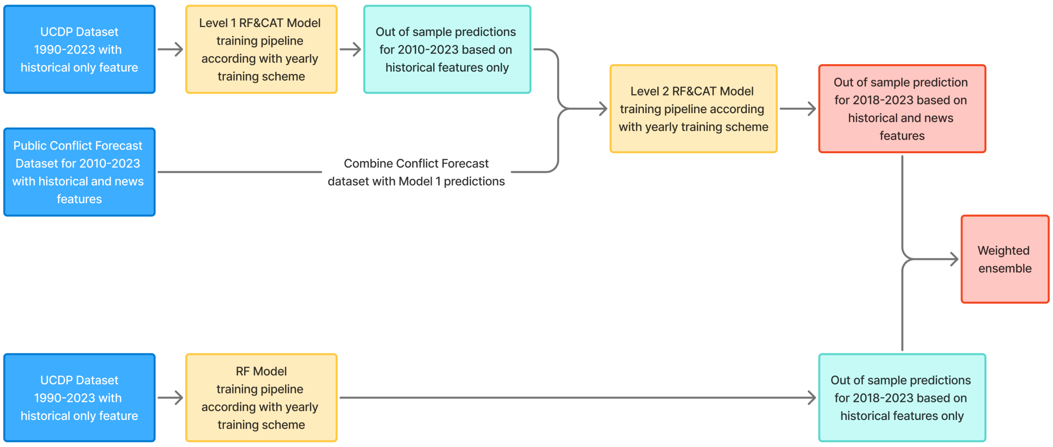

UCDP Dataset 1990-2023 with historical only feature
Level 1 RF&CAT Model training pipeline according with yearly training scheme
Out of sample predictions for 2010-2023 based on historical features only
Public Conflict Forecast Dataset for 2010-2023 with historical and news features
Combine Conflict Forecast dataset with Model 1 predictions
Level 2 RF&CAT Model training pipeline according with yearly training scheme
Out of sample prediction for 2018-2023 based on historical and news features
Weighted ensemble
UCDP Dataset 1990-2023 with historical only feature
RF Model training pipeline according with yearly training scheme
Out of sample predictions for 2018-2023 based on historical features only

Figure 5.3: Combined stacking and weighted ensembling approach. The predictions of the stacking pipeline (UCDP historical features enriched with Conflict Forecast's news-based predictions) and the separate UCDP historical-only model are combined through optimized weighted ensembling to yield improved predictive performance.

topic modeling to summarize news data into fixed topics [15], we attempt to directly encode the entire textual content of monthly summaries into vectors. The main advantage of this approach is the retention of detailed textual information rather than reducing it solely to broad topics. The reasoning for such approach is that with vectorised text, the model can learn more nuanced relationships between the monthly summary of what has happened in the country, structural variables and the armed conflict risk, instead of relying only on slowly moving structural variables.

To construct these news features, we scrape concise yet detailed monthly summaries from the CrisisWatch website [39]. This dataset spans approximately 21 years (2003–2024), covering around 80 countries per month, resulting in roughly 20,000 country-month observations. Although valuable, this dataset size is fairly small for training multimodal BERT-based deep-learning models [40].

Figure 5.4 illustrates the pipeline we follow in this experiment. We first encode the textual summaries into numeric vectors using AutoGluon's multimodal package, which natively supports combining textual and numeric features [41]. These encoded vectors are then merged with structural variables from the UCDP dataset. The combined dataset is used to train a multimodal model that predicts armed conflict risk.

Despite the conceptual appeal of this approach, our experiments did not yield significant improvements. The multimodal model struggled to learn meaningful relationships from the data. This outcome is likely due to the limited dataset size—approximately 20,000 samples—being insufficient to train a deep-learning model with such high-dimensional textual data effectively. Future studies could explore larger datasets or alternative methods, such as transfer learning from pretrained language models, to better leverage textual information. A great attempt to implement such approach that does not lose information was done by the ViEWS team using Large Language Models and indiviadual news [20].

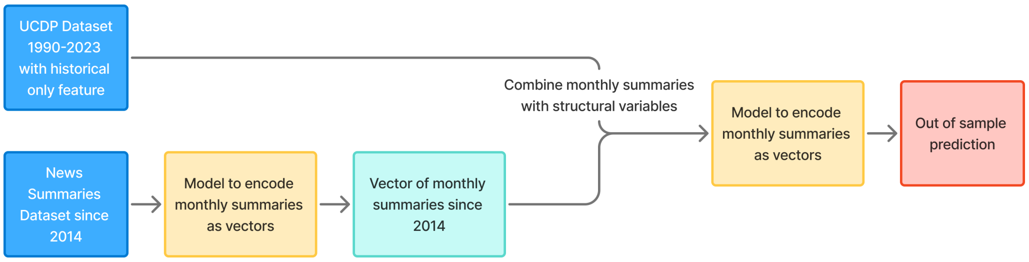

UCDP Dataset 1990-2023 with historical only feature
News Summaries Dataset since 2014
Model to encode monthly summaries as vectors
Vector of monthly summaries since 2014
Combine monthly summaries with structural variables
Model to encode monthly summaries as vectors
Out of sample prediction

Figure 5.4: News modelling pipeline that vectorizes the monthly CrisisWatch summaries for each country, combines them with the structural features, and trains a model to predict the risk of armed conflict.

6
Results

In this section, we present the results of our experiments on the three datasets: ViEWS, Conflict Forecast, and UCDP. We evaluate the three machine learning approaches described in Section 4.3, as well as for our improvements described in Section 5.

When comparing models and analysing the results, we focus on the Precision–Recall AUC (PR-AUC) score as our primary metric. Nevertheless, we also report the Area Under the Receiver Operating Characteristic (ROC-AUC) score for completeness in the Appendix Section B

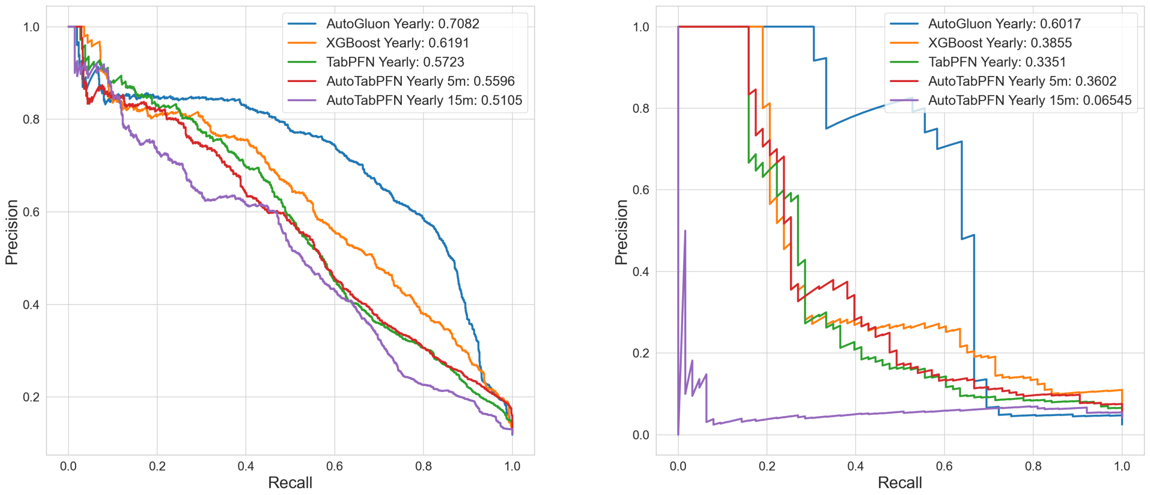
6.1. Model Performance on ViEWS data

In this section, we present the results of running our three main models—AutoGluon, XGBoost, and TabPFN—as well as two variations of AutoTabPFN, on the ViEWS dataset. We use the yearly training scheme described earlier, meaning we retrain each model once per forecast year. Additionally, for AutoTabPFN, we ran models with computational budgets of 5 and 15 minutes to see their relative performance.

We find that AutoGluon clearly outperforms all other models when predicting all conflict onsets, achieving a Precision–Recall AUC (PR-AUC) of 0.708. This result is substantially better compared to XGBoost (0.619) and the vanilla TabPFN model (0.572). Interestingly, while AutoTabPFN with a 5-minute computational budget (PR-AUC = 0.560) slightly underperforms compared to vanilla TabPFN for all onsets, AutoTabPFN with a 15-minute budget performs even worse (PR-AUC = 0.511). This unexpected result aligns with the empirical tests we performed later, which showed that AutoTabPFN only starts outperforming the vanilla TabPFN model after approximately 22 minutes of computation.

Predicting *hard onsets*, defined as conflicts emerging after at least five years of peace (see Section 4.2), is significantly more challenging for all models We see that AutoGluon achieves the highest PR-AUC of 0.363. The vanilla TabPFN significantly drops in performance (PR-AUC = 0.127), while AutoTabPFN with a 5-minute budget slightly improves over vanilla TabPFN, achieving a PR-AUC of 0.144. Again, the AutoTabPFN with a 15-minute budget performs poorly, obtaining a PR-AUC of only 0.044.

Overall, these results highlight AutoGluon's consistent strength in predicting conflicts using the ViEWS dataset. The findings regarding AutoTabPFN's relative underperformance at 5 and 15-minute computational budgets are consistent with our later observations described in Section 4.3.4, indicating that AutoTabPFN requires a longer runtime to surpass vanilla TabPFN performance. Additionally, due to the large dataset size, TabPFN cannot perform optimally. These results are visualized in Figure 6.1a and Figure 6.1b, which show the PR-AUC performance of all models from 2018 to 2023 for all onsets and hard onsets respectively. It's also important to mention that these ViEWS results cannot be directly compared with Conflict Forecast and UCDP datasets, because although the dependent variable is calculated similarly, the underlying fatality counts differ, making the results inherently incomparable.



(a) Model performance on ViEWS dataset (2018–2023), all onsets.

(b) PR-AUC accuracy of our models on ViEWS dataset (2018–2023), hard onsets.

Figure 6.1: Precision–Recall AUC score comparison of models on the ViEWS dataset for all onsets and hard onsets.

6.2. Model Performance on Conflict Forecast Data

In this section, we present the performance results of the various models trained on the Conflict Forecast dataset over the period from 2018 to 2023. To benchmark our models we use the best three model published by Conflict Forecast team: the text-only model (CF Text), the historical-features model (CF Hist), and the full-feature model (CF All). Figure 6.2a and Figure 6.2b illustrate the Precision–Recall AUC (PR-AUC) scores for all our models across all conflict onsets and hard onsets and how they compare to the best-performing Conflict Forecast models.

Analysing the performance across all conflict onsets, we observe an intriguing set of results. The AutoGluon yearly model, trained using all available models without hyperparameter optimisation (exactly as done for the ViEWS dataset), achieves the lowest PR-AUC of 0.6528. This indicates that despite its strength on the ViEWS dataset, AutoGluon does not perform as well on Conflict Forecast data under these settings. This outcome is unexpected given the model's generally robust performance in other experiments and highlights the potential impact of model selection and the absence of hyperparameter tuning, which will be explored further in the subsequent section on UCDP dataset analysis.

The vanilla TabPFN model significantly outperforms AutoGluon, demonstrating an effective capacity to capture the complexities within this smaller dataset. However, the AutoTabPFN ensemble trained with a yearly scheme over 1 hour shows a slightly worse performance compared to the simpler TabPFN model. This is contrary to our expectations, as the additional computational resources dedicated to AutoTabPFN were shown to typically improves model performance [37]. The Conflict Forecast dataset is three times larger than the recommended size for TabPFN and this likely causes model performance degradation, despite the fact that both TabPFN and AutoTabPFN, do support larger datasets [42].

The best-performing yearly trained model is XGBoost, achieving a PR-AUC that surpasses both AutoGluon and the TabPFN yearly models. Yet, notably, implementing a monthly optimisation scheme for XGBoost does not yield any substantial performance improvement, suggesting limited incremental gains from continuous monthly updates for this particular model on the Conflict Forecast dataset.

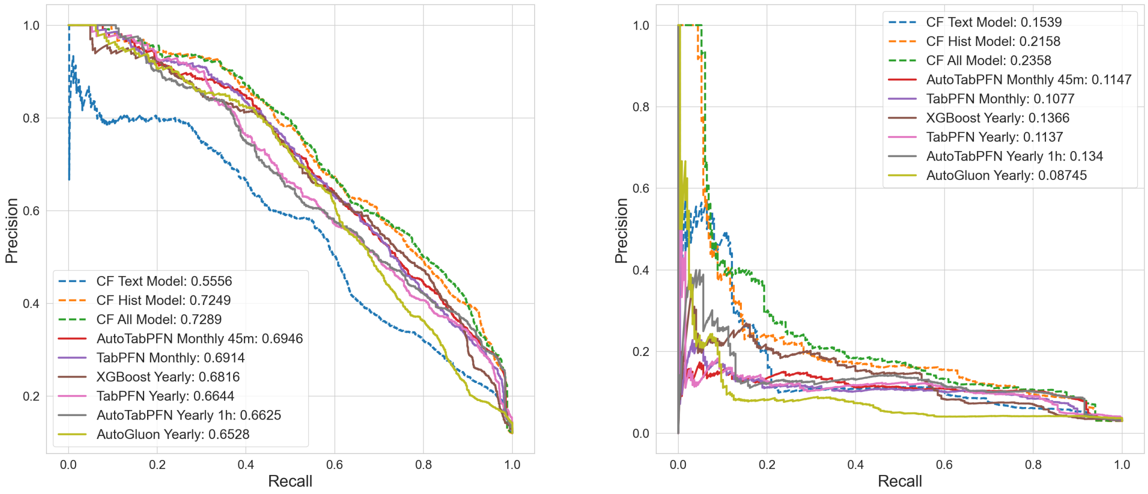
Encouraged by the promising results of the TabPFN model, we conducted a monthly optimisation scheme for this model, significantly enhancing its performance. This monthly optimisation of TabPFN yields a PR-AUC of 0.6914, representing a meaningful increase of 4.06% over the yearly-trained TabPFN model, which has a PR-AUC of 0.6644.

To maximise the potential of the AutoTabPFN model on the Conflict Forecast dataset, we run an AutoTabPFN monthly training scheme using a computational budget of 45 minutes per month, leveraging Nvidia L40S GPUs for a total of 54 hours of computation. The resulting model achieved the highest PR-AUC of 0.6946, surpassing all other models tested on this dataset. However, this result is still not enough to catch up with the best Conflict Forecast model, which has a PR-AUC of 0.7289.

Focusing specifically on hard onsets, all models experience a notable decrease in predictive performance, underscoring the inherent challenge in forecasting outbreaks in countries with prolonged peace, even with the news LDA topic modelling. Our best AutoTabPFN monthly model performs twice as poorly as the best model of Conflict Forecast on hard onsets.

Overall, we see that even our best-performing model on the Conflict Forecast dataset is still considerably behind the top Conflict Forecast models for both all and hard onsets. Our assumption is that while the best Conflict Forecast models are trained on the full dataset, the limited dataset that is available to the public is not enough for the models to effectively learn and generalise.

Therefore, we decided to replicate the Conflict Forecast dataset by constructing our own version based on the UCDP data (see the Section 6.3 for performance on this dataset and Section 3.3 for details on how we constructed the dataset). The results of our models trained on this newly created dataset are presented in the next subsection.



(a) Model performance on Conflict Forecast dataset (2018–2023), all onsets.

(b) Model performance on Conflict Forecast dataset (2018–2023), hard onsets.

Figure 6.2: Precision–Recall AUC score comparison of models on the Conflict Forecast dataset for all onsets and hard onsets.

6.3. Model Performance on UCDP data

In this section, we evaluate how well our models perform on the replicated UCDP dataset for the years 2018 to 2023. Our goal throughout this study has been to match and ideally surpass the performance of the Conflict Forecast team. Figure 6.3a and Figure 6.3b report the performance of our models compared to the best Conflict Forecast models.

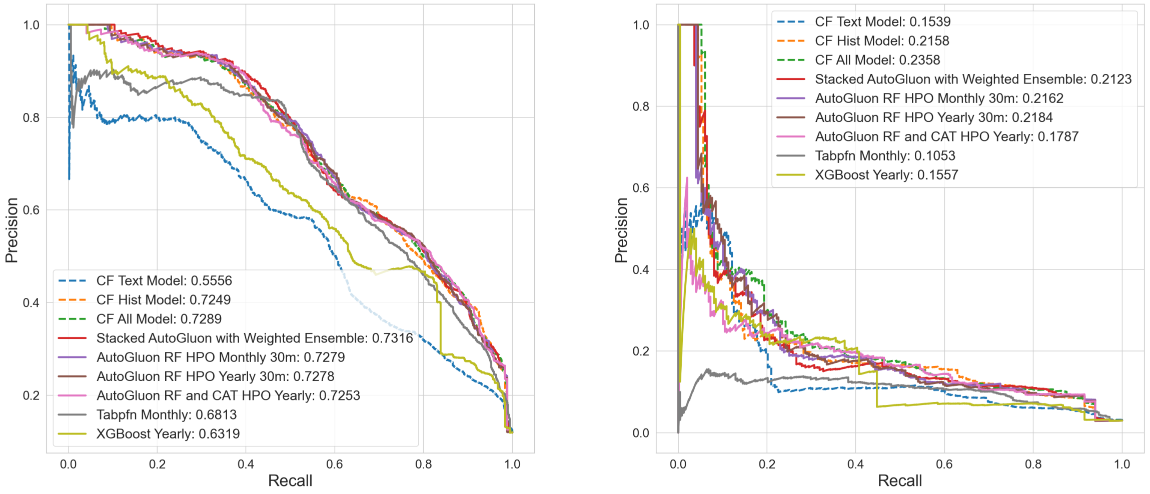
When looking at the prediction of all conflict onsets, our models show promising results. Our simpler approaches, such as XGBoost Yearly (PR-AUC = 0.6319) and TabPFN Monthly (PR-AUC = 0.6813), perform reasonably well but still fall short compared to the CF models, especially the CF All model (PR-AUC = 0.7289). However, once we move to more advanced methods like AutoGluon with yearly hyperparameter optimisation (PR-AUC = 0.7278) and monthly hyperparameter optimization (PR-AUC = 0.7279), our models quickly catch up, nearly matching CF's best model performance.

The finding that TabPFN as well as AutoTabPFN models do not perform well on large datasets is consistent with TabArena benchmark that highlights that Tabular foundational models are still not ready for large datasets [42]. At the same time, poor XGBoost performance likely means that the selected range of the hyperparameters is suboptimal for the UCDP dataset, as the same hyperparameters range on the Conflict Forecast dataset yielded better results (see Section 6.2).

Most importantly, our stacked AutoGluon model, which combines multiple learners, reaches a PR-AUC of 0.7316. This result not only successfully reproduces the performance of the CF All model but slightly exceeds it, marking a significant achievement and highlighting the value of carefully combining diverse models.

The story changes when predicting hard onsets. While CF All still holds the best performance with a PR-AUC of 0.2358, our best model—the AutoGluon Random Forest with yearly optimisation—achieves a PR-AUC of 0.2184. Our stacked AutoGluon ensemble achieves 0.2123, clearly surpassing the simpler CF Text model (0.1539) but still lagging behind CF All and CF Hist.

The results on the UCDP dataset demonstrate that we have successfully reproduced the Conflict Forecast models' performance. By leveraging advanced methods like AutoGluon and stacking different models, we even managed to slightly improve upon the original CF benchmark for all onsets. However, the hard-onset predictions remain challenging.



(a) Model performance on UCDP dataset (2018–2023), all onsets.

(b) Model performance on UCDP dataset (2018–2023), hard onsets.

Figure 6.3: Precision–Recall AUC score comparison of models on the UCDP dataset for all onsets and hard onsets.

6.4. Predictions into the real future

This section presents the predictions into the real future of our best model Stacked Weighted ensemble model trained UCDP dataset. The model predicts the maximum risk of armed conflict in the period from 2025–05 to 2026–04. To produce predictions into the real future, we train the model of the data from 1990 to 2024–04 and predict the risk of armed conflict based on the 2024–05, which was the latest month available at the time of analysis. Figure 6.4 depicts the risk predictions of our best model for covered country.

Analysing the predictions, we see that Sub-Saharan Africa is identified as the Epicenter of Predicted Conflict. The model identifies several African nations as facing extreme conflict risk, with Ethiopia leading at 0.911 risk probability Recent developments in Ethiopia's Tigray region validate this prediction, as renewed fighting between the Tigray People's Liberation Front and federal forces threatens to escalate into a broader regional war. The incomplete

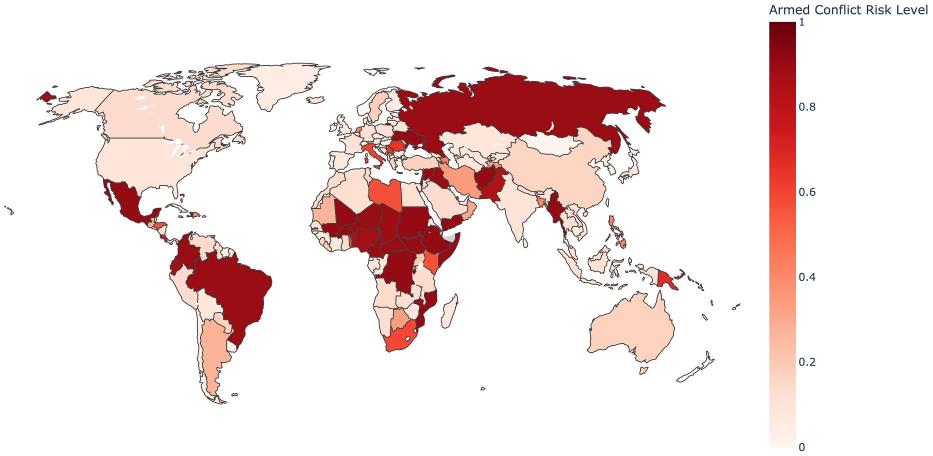


Figure 6.4: Our best model predictions of the armed conflict risk for covered countries based on inputs as of May 2025.

implementation of the 2022 Pretoria Agreement has created power vacuums that both Eritrea and Ethiopia are positioning to exploit.

Mali emerges as the second-highest risk country with a 0.910 probability, reflecting the devastating reality of ongoing conflict since 2012. The country faces simultaneous challenges from jihadist insurgencies, Tuareg rebellions, and the presence of foreign mercenary forces, creating a complex multi-front conflict that has already displaced over two million people. Recent attacks in 2024 have demonstrated the escalating brutality of the conflict, with civilian casualties reaching unprecedented levels .

Burkina Faso, with a 0.910 risk score, has become the epicenter of armed group violence in the Central Sahel region. The country now ranks first in the Global Terrorism Index, accounting for nearly a quarter of all terrorist deaths globally. Armed groups control approximately half of the national territory, implementing devastating blockades that have affected over one million civilians across 46 locations[43].

Predictions for Europe and Asia regions also align with the current geopolitical landscape. The prediction for Ukraine is 0.91, which reflect the continuation of the devastating conflict that began with Russia's 2022 invasion. Syria maintains a 0.906 risk level, indicating that despite reduced international attention, the underlying conditions for renewed large-scale conflict remain present . The country's fractured state and presence of multiple foreign powers continue to create instability that could rapidly escalate. Myanmar's 0.908 risk score accurately reflects the country's descent into civil war following the 2021 military coup. The conflict has killed over 50,000 people and displaced approximately three million, with resistance forces gaining significant territorial control throughout 2024.

The model also identifies several Middle Eastern countries in high-risk categories. Afghanistan's 0.907 score reflects ongoing instability under Taliban rule, while Iraq's 0.898 rating indicates persistent security challenges despite reduced international presence. At the same time, Somalia's inclusion with a 0.908 risk probability demonstrates the continued threat posed by al-Shabaab and the complex regional dynamics involving Ethiopian intervention.

7 Discussion

Currently, fully open-source academic models for early conflict forecasting are two years behind state-of-the-art closed-source models. This work demonstrates that, despite this gap, it is possible to catch up with and even slightly surpass current state-of-the-art models by relying entirely on open-source tools, such as the AutoGluon framework, combined with careful stacking and ensembling methods. This finding confirms the effectiveness of ensemble techniques for conflict prediction, particularly given the unique challenges of this domain.

One critical insight from this research is the importance of careful modeling strategies due to the nature of early conflict forecasting data. Conflict events are rare and highly imbalanced, making accurate predictions especially challenging. A robust validation strategy, attentive hyperparameter optimization, and careful handling of the data to avoid information leaks proved essential to achieving strong predictive results.

However, significant challenges remain. Predicting the *hard onsets* (conflicts emerging in historically stable countries after prolonged periods of peace), continues to be difficult. All structural models struggle with accurately forecasting these rare events as historical patterns offer limited predictive power. Recent promising attempts, such as the approach described by Croicu and von der Maase [20], suggest potential for big improvements in combining textual and structural data. The goal of such approach is to improve understanding of conflict dynamics to better forecast hard onsets. However, these efforts currently still face limitations related to data quality and computational resources. More high-quality, multimodal datasets and refined methods for combining text-based and structural data are needed.

For policymakers, this work provides evidence that practical, data-driven forecasting models for conflict are achievable. Yet, these models remain imperfect and carry inherent risks. Predicting conflicts that never occur, so-called false positives, could lead to misallocated resources or diplomatic tensions, while failing to predict an emerging conflict (false negatives) might allow crises to escalate unnoticed. Consequently, predictions from early conflict warning models should always be paired with human verification and expert analysis before informing critical policy decisions.

For researchers interested in early conflict modeling, this work provides an openly available framework, datasets, and detailed pipelines. By openly sharing the developed tools and code, we encourage further research and refinement. Our hope is that the availability of these resources can help the research community accelerate advancements in early conflict forecasting, promoting greater transparency, reproducibility, and collaboration in the field.

7.1. Ethical Considerations

From an ethical perspective, the potential societal benefits of early conflict warning are substantial. Reliable, timely forecasts can help direct scarce preventive resources to places where they are most likely to prevent loss of life, improve accountability in prioritisation, and make the underlying trade-offs between false positives and false negatives explicit to decision-makers [15]. In practice, embedding model outputs in transparent decision frameworks—complete with calibrated thresholds, clear uncertainty communication, and routine validation can support earlier diplomatic engagement and targeted, less intrusive preventive action [25]. Open datasets and reproducible pipelines also enable independent audit and continuous improvement of models by a wider research and policy community, which is morally right when forecasts can shape life-or-death decisions [16].

At the same time, forecasting tools are not neutral: they operate in reflexive, strategic environments where information can alter behaviour. Publicly labelling countries or subregions as “high risk” can discredit communities, shift capital and aid in harmful ways, or, under certain political conditions, be invoked to justify pre-emptive repression. Text-driven features inherit biases from news ecosystems (for example, censorship or uneven coverage), which can skew risk assessments and distributive choices if left unexamined [16]. These concerns reinforce long-standing cautions from the EWS field about “do no harm” practices, user-centred design, and tight coupling between warnings and appropriate, proportionate responses [25]. Accordingly, we view open production forecasting as ethically defensible only with safeguards: publish uncertainty and model limits alongside any scores; audit for bias and drift; document data sources; restrict highly granular outputs when they could increase targeting risks; and establish governance for responsible use (including red-teaming and after-action reviews) so that forecasts inform prevention rather than inadvertently escalating conflict dynamics [15], [44], [45].

8

Conclusion & Future Work

In this work, we applied and evaluated multiple machine learning models applied to the early conflict forecasting domain and caught up the open source models with the Conflict Forecast state-of-the-art models without having access to the full dataset with the news topics. In order to match the performance of state-of-the-art models, we built a robust, reproducible and extensible framework for early conflict prediction, as well as our custom pipelines to build datasets and evaluate models. Our goal was to contribute to a future where conflicts can be reliably anticipated and proactively mitigated with the improved early conflict warning systems.

We achieved five key contributions:

- We developed a flexible framework for Early Conflict Modelling, enabling streamlined training, evaluation, and benchmarking of predictive models.
- We implemented a training and evaluation pipelines for state-of-the-art machine learning models applied to the task of early conflict forecasting. We covered diverse range of models starting from tree-based Random Forest models, and including an automated machine learning framework and a tabular foundation model.
- We reproduced the historical dataset pipeline used by the Conflict Forecast team, allowing for fair comparisons and further reproducibility in the field.
- We built a model that not only catches up with but slightly outperforms the best existing Conflict Forecast baseline, demonstrating competitive results for country-level forecasts.
- We tested multiple approaches and we reported unsuccessful attempts to combine text data with structural data to allow future researches to reuse our news datasets and pipelines.

Despite the advancements, our findings underscore the limitations of purely structural and tabular machine learning approaches when tackling the complex problem of armed conflict prediction. While such models offer valuable signals, they often fall short in capturing the nuanced socio-political dynamics and fast-evolving realities that precede conflict. This leads to limited accuracy on, arguably, the most valuable aspect of proactive conflict prediction - emergence of conflict in countries with a long history of peace (so called hard onsets).

We believe the next major leap in early conflict forecasting will come from integrating structural data with the contextual intelligence of large language models. These models offer the capability to ingest and interpret high-quality political research, historical relationships, and real-time geopolitical developments present in news and policy documents. LLMs can grasp latent factors such as diplomatic tensions, alliances, rhetoric, and public sentiment—elements that are difficult to encode explicitly in tabular data.

Our future work will focus on building a hybrid forecasting architecture that combines these strengths: the quantitative rigor of structured indicators and the qualitative depth of natural language understanding. By linking LLM-derived insights with structural predictors, we aim to enable earlier, more precise, and context-aware conflict forecasts. We believe such systems are a crucial step toward proactive peacebuilding and, eventually, global stability.

References

[1] S. Davies, T. Pettersson, M. Sollenberg, and M. Öberg, "Organized violence 1989–2024, and the challenges of identifying civilian victims," *Journal of Peace Research*, vol. 62, no. 4, 2025.

[2] R. Sundberg and E. Melander, "Introducing the ucdp georeferenced event dataset," *Journal of Peace Research*, vol. 50, no. 4, pp. 523–532, 2013.

[3] DefenceWeb, "Global conflicts doubled over the past five years, set to worsen in 2025," *DefenceWeb*, 2024, Accessed: 2025-07-19. [Online]. Available: https://www.defenceweb.co.za/joint/diplomacy-a-peace/global-conflicts-doubled-over-the-past-five-years-set-to-worsen-in-2025/

[4] H. Mueller, C. Rauh, B. R. Seimon, and R. A. Espinoza, "The urgency of conflict prevention: A macroeconomic perspective," International Monetary Fund, Tech. Rep. WP/25/43, 2025, Accessed: 2025-07-19. [Online]. Available: https://www.imf.org/en/Publications/WP/Issues/2024/12/17/The-Urgency-of-Conflict-Prevention-A-Macroeconomic-Perspective-559143

[5] A. Austin, M. Fischer, and N. Ropers, *Advancing conflict transformation: The berghof handbook ii*, Accessed: 2025-07-13, 2011. [Online]. Available: https://berghof-foundation.org/files/publications/austin_handbook.pdf

[6] H. Hegre and E. G. Rød, "A review and comparison of conflict early warning systems," *PRIO Paper*, 2022.

[7] G. King and L. Zeng, "State failure task force report: Phase iii findings," *CIDCM Working Paper*, 2000.

[8] J. A. Goldstone et al., "A global model for forecasting political instability," *American Journal of Political Science*, vol. 54, no. 1, pp. 190–208, 2010.

[9] M. D. Ward, B. D. Greenhill, and K. M. Bakke, "Forecasting civil conflicts with machine learning: Pitfalls and promises," *International Studies Quarterly*, vol. 57, no. 2, pp. 472–483, 2013.

[10] M. Colaresi and Z. Mahmood, "Forecasting at the segment level in international relations: A machine learning approach to forecasting conflict," *Conflict Management and Peace Science*, vol. 34, no. 1, pp. 99–117, 2017.

[11] H. Hegre, L. Hultman, H. M. Nygård, J. Rydgren, et al., "Views: A political violence early-warning system," *Journal of Peace Research*, vol. 56, no. 2, pp. 155–174, 2019. DOI: 10.1177/0022343319823860

[12] The ViEWS team, "Forecasting fatalities," Department of Peace, Conflict Research, Uppsala University, and Peace Research Institute Oslo (PRIO), Tech. Rep., May 2022, Funded by UK aid from the UK government.

[13] V. Team, *Views prediction challenge 2023/24*, urlhttps://viewsforecasting.org/prediction-challenge-2023-24/, Accessed: 2025-07-13, 2024.

[14] A. Carlson et al., "Views fatalities prediction competition: Methodology and outcomes," *Unpublished manuscript*, 2023.

[15] H. Mueller and C. Rauh, "The hard problem of prediction for conflict prevention," *Journal of the European Economic Association*, vol. 20, no. 6, pp. 2440–2467, 2022. DOI: 10.1093/jeea/jvac025

[16] H. Mueller, J. Rauh, et al., "Introducing a global dataset on conflict forecasts and news topics," *Data Policy*, vol. 6, e16, 2024. DOI: 10.1017/dap.2024.16

[17] H. Mueller et al., "Using past violence and current news to predict changes in violence," *Political Science Research and Methods*, 2024.

[18] P. T. Brandt et al., *Conflibert: A language model for political conflict*, 2024. arXiv: 2412.15060 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.15060

[19] V. d'Orazio et al., "Eventbert: Fine-tuning language models to detect political events," *arXiv preprint arXiv:2403.00000*, 2024.

[20] M. Croicu and S. von der Maase, "From newswire to nexus: Forecasting dyadic conflict using transformer embeddings," *Working paper*, 2025.

[21] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward, "Icews coded event data," *Harvard Dataverse*, 2015.

[22] C. Raleigh et al., "Acled: Armed conflict location & event data project," *Journal of Peace Research*, 2021, ACLED overview publication.

[23] N. Marco and P. Ball, "Statistical risk assessment: Early warning project," *Journal of Genocide Research*, vol. 23, no. 2, pp. 212–229, 2021.

[24] M. Halkia, S. Ferri, M. K. Schellens, M. Papazoglou, and D. Thomakos, "The global conflict risk index. a quantitative tool for policy support on conflict prevention," *European Journal of Political Research*, vol. 59, no. 4, pp. 939–961, 2020.

[25] T. Lynam, M. Zapata, H. Hegre, C. Bell, and C. Besaw, "Early warning and predictive analytic systems in conflict contexts: Insights from the field," *Civil Wars*, vol. 26, no. 3, pp. 401–429, 2024. DOI: 10.1080/13698249.2023.2185377

[26] *The World Bank Annual Report 2015*. The World Bank, 2015. DOI: 10.1596/978-1-4648-0574-5

[27] M. Coppedge et al., "Conceptualizing and measuring democracy: A new approach," *Perspectives on Politics*, vol. 9, no. 2, pp. 247–267, 2011. DOI: 10.1017/s1537592711000880

[28] L.-E. Cederman, A. Wimmer, and B. Min, "Why do ethnic groups rebel? new data and analysis," *World Politics*, vol. 62, no. 1, pp. 87–119, 2009. DOI: 10.1017/s0043887109990219

[29] W. Lutz, "Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000," *Vienna Yearbook of Population Research*, vol. 2007, pp. 193–235, 2007. DOI: 10.1553/populationyearbook2007s193

[30] C. Raleigh, r. Linke, H. Hegre, and J. Karlsen, "Introducing acled: An armed conflict location and event dataset," *Journal of Peace Research*, vol. 47, no. 5, pp. 651–660, 2010. DOI: 10.1177/0022343310378914

[31] Food and A. Organization, *Aquastat glossary*, https://www.fao.org/aquastat/en/, FAO website, 2019.

[32] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," pp. 785–794, 2016.

[33] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," pp. 2623–2631, 2019.

[34] N. Erickson et al., "Autogluon-tabular: Robust and accurate automl for structured data," *arXiv preprint arXiv:2003.06505*, 2020.

[35] N. Hollmann et al., “Accurate predictions on small data with a tabular foundation model,” *Nature*, vol. 637, no. 8045, pp. 319–326, 2025.

[36] N. Hollmann et al., “Foundation models for tabular data outperform specialized algorithms and transfer across tasks,” *Nature*, vol. 630, no. 8039, pp. 756–763, 2024. DOI: `10.1038/s41586-024-08328-6`

[37] H.-J. Ye, S.-Y. Liu, and W.-L. Chao, “A closer look at tabpfn v2: Understanding its strengths and extending its capabilities,” *arXiv Preprint*, vol. arXiv:2502.17361v2, 2025. [Online]. Available: `https://arxiv.org/abs/2502.17361v2`

[38] e. a. Clayton Besaw PhD, “Annual risk of coup report,” One Earth Future, Tech. Rep., 2019. DOI: `10.18289/OEF.2019.037` [Online]. Available: `https://oneearthfuture.org/sites/default/files/documents/publications/Risk_of_Coup_Report_2019_0.pdf`

[39] International Crisis Group, *Crisiswatch monthly summaries*, Accessed: 2025-07-19, 2024. [Online]. Available: `https://www.crisisgroup.org/crisiswatch`

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: `https://arxiv.org/abs/1810.04805`

[41] Z. Tang et al., “Autogluon-multimodal (automm): Supercharging multimodal automl with foundation models,” *arXiv preprint arXiv:2404.16233*, 2024. [Online]. Available: `https://arxiv.org/abs/2404.16233`

[42] N. Erickson et al., “Tabarena: A living benchmark for machine learning on tabular data,” *arXiv preprint arXiv:2506.16791*, 2025.

[43] ACAPS, *Burkina faso*, Accessed: 2025-06-21, 2025. [Online]. Available: `https://www.acaps.org/en/countries/burkina-faso`

[44] S. Bazzi, R. A. Blair, C. Blattman, O. Dube, M. Gudgeon, and R. Peck, “The promise and pitfalls of conflict prediction: Evidence from colombia and indonesia,” *The Review of Economics and Statistics*, vol. 104, no. 4, pp. 764–779, 2022. DOI: `10.1162/rest_a_01016` [Online]. Available: `https://doi.org/10.1162/rest_a_01016`

[45] W. Guo, K. S. Gleditsch, and A. Wilson, “Retool AI to forecast and limit wars,” *Nature*, vol. 562, no. 7727, pp. 331–333, 2018. DOI: `10.1038/d41586-018-07026-4` [Online]. Available: `https://www.nature.com/articles/d41586-018-07026-4`

A

UCDP Pipeline Implementation details

This appendix outlines the full data engineering pipeline developed to reconstruct the Conflict Forecast historical dataset using publicly available UCDP data. The pipeline consists of 16 sequential steps, each transforming and enriching the dataset toward a structure suitable for early conflict modeling. Each step is a separate implementation in Python or R, available to the reader. Below we describe each step in detail:

Step 1: Download Raw UCDP Data

We begin by programmatically downloading the most recent UCDP GED (Georeferenced Event Dataset) and Candidate GED files. These files include finalized and provisional conflict events. All downloads are stored locally for consistent preprocessing.

Step 2: Add Lacking Countries

Certain countries are missing from the UCDP source but exist in Conflict Forecast with all 0 fatalities. To understand which countries should be added for for which periods, we merge in conflict rows from latest Conflict Forecast data for post-2010 months and the 2022 Conflict Forecast dataset for pre 2010 months, allowing us to reconstruct such countries.

Step 3: Fix Country Time Series Lengths

Some countries in UCDP dataset may have fatalities recorded earlier or later than the country starts existing according to Conflict Forecast definitions. To remain consistent with Conflict Forecast, we adjust country time series by tracking which months are included for which countries in the original Conflict Forecast datasets. For example, countries like South Sudan, Kosovo, and others that were not existent in 1989 are set to be excluded up until their first appearance in either UCDP or Conflict Forecast data.

Step 4: Fix Fatality Values

This step corrects known anomalies in the fatality data. For instance, we set all fatalities for Serbia to zero, following the behavior observed in Conflict Forecast preprocessing. These corrections ensure the consistency of fatality values across the datasets for some selected problematic countries.

Step 5: Add Population Data

Monthly population estimates are constructed by merging World Bank data (1989–2009) with our dataset. Missing countries and years are filled by interpolation and backfilling, ensuring population coverage across the entire time span.

Step 6: Assign COW Country Codes

We map ISO3C country codes to their corresponding COW (Correlates of War) numeric codes. Manual corrections are applied for countries with inconsistent mappings. These codes are used later to generate neighbor-based features using geographic proximity.

Step 7: Compute Armed Conflict and Discounted Features

We compute binary indicators for each country-month: `armedconf` (if fatalities per million exceed 0.5), and `anyviolence` (if any fatalities are recorded). Exponentially discounted versions of these indicators are calculated using a decay factor of 0.95, to match `discounted_*` variables in Conflict Forecast.

Step 8: Generate Neighbor-Based Features

Using the Correlates of War minimum distance matrix, we compute regional spillover features. These include neighbor-averaged versions of the discounted conflict variables. For each country, we calculate the average of `discounted_armedconf` and `discounted_anyviolence` for all countries that share a border (minimum distance is exactly 0). In case a country has no neighbors, the feature is set to zero. Neighboring dynamics are a key component in identifying contagion effects in regional violence.

We introduce some corrections to the Correlates of War minimum distance matrix to be able to precisely follow the unpublished distance matrix used in Conflict Forecast. These corrections are stored in a separate configurations file.

Step 9: Civil War Indicator

We define a civil war as any month in which fatalities exceed 0.003% of a country's population. This binary `civilwar` variable helps isolate high-intensity conflict months and serves as a basis for several temporal indicators in future steps.

Step 10: Ongoing Conflict Features

We compute ongoing feature for each binary conflict type (`anyviolence`, `armedconf`, `civilwar`) that implemented as a counter that record the number of consecutive months the country has experienced that type of conflict. These counters are reset when the conflict ends.

Step 11: Months Since Last Conflict

Complementing the ongoing counters, we calculate the number of months since each type of conflict last occurred. These *since* features provide a memory of peaceful periods. Manual adjustments are applied for edge cases where countries were historically excluded or underrepresented.

Step 12: Rolling Fatalities Features

We compute rolling sums of fatalities per capita over multiple past windows (e.g., 3, 6, 12, 24 months), scaled per 1,000 people. These "`past_bestpc`" features quantify recent violence intensity.

Step 13: Period Identifier

To simplify monthly indexing and ensure alignment with external datasets, we create a `period` variable in YYYY–MM format that precisely follow the period definitions used in Conflict Forecast.

Step 14: Target Variable

We define the dependent variable (`ons_armedconf_12_target`) as an indicator for conflict onset within the next 12 months, provided the current month is peaceful. The full definition is described in Section 4.1. If a conflict is already ongoing, the target is masked as missing. Additionally, the final 12 months of data are masked to prevent future leakage during model training.

Step 15: Attach Conflict Forecast Predictions

To enable benchmarking, we merge in historical predictions from the Conflict Forecast models, including their text-only, historical-only, and full-feature versions. These predictions are used for validation and performance comparisons throughout the thesis.

Step 16: Select Final Columns

Finally, we select the subset of variables defined as historical features. Intermediate columns and temporary features are dropped, and the final processed dataset is saved. This cleaned dataset serves as the basis for all model training and evaluation in this work.

B

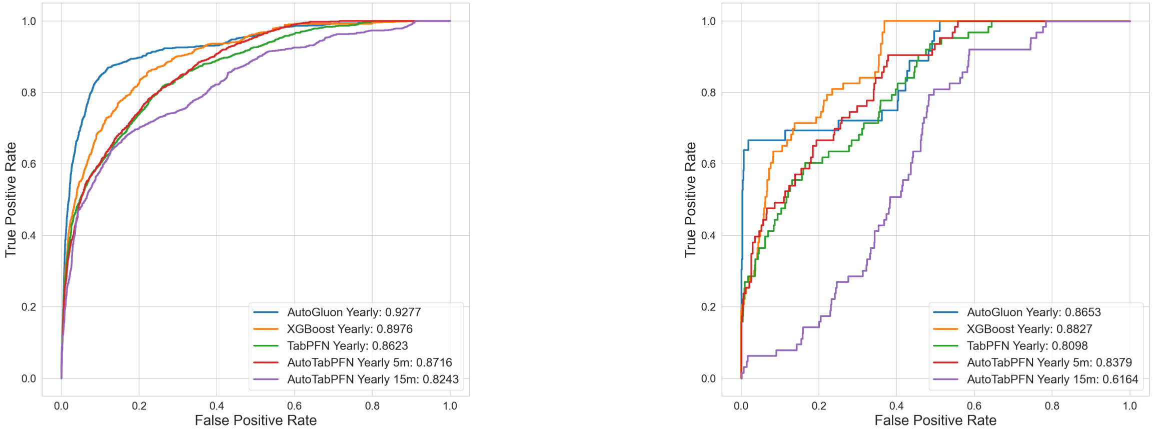
ROC-AUC Scores for all models

This appendix provides the detailed Receiver-Operator AUC (ROC-AUC) scores for all models evaluated in this work for the three datasets: ViEWS, Conflict Forecast, and UCDP. As described in the Section 3 and Section 4.1, ROC-AUC is not a representative metric for the imbalanced datasets, but we still report it for the reproducibility and benchmarking reference.

Analysing the ROC-AUC scores, we see that the results are generally inline with the Precision-Recall AUC scores reported in the Section 6 with minor differences. Our best model still dominates the ROC-AUC scores, which is in line with the expectations.

B.1. ViEWS Dataset ROC-AUC Scores

This subsection in the Figure B.1 presents the ROC-AUC scores for all models evaluated on the ViEWS dataset from 2018 to 2023.



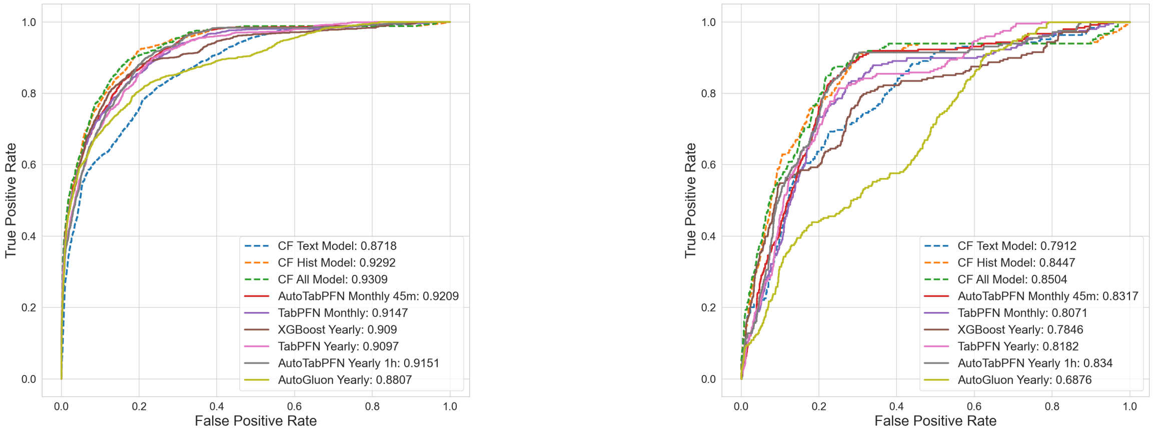
(a) Model performance on ViEWS dataset (2018–2023), all onsets.

(b) Model performance on ViEWS dataset (2018–2023), hard onsets.

Figure B.1: ROC-AUC score comparison of models on the ViEWS dataset for all onsets and hard onsets.

B.2. Conflict Forecast Dataset ROC-AUC Scores

This subsection in the Figure B.2 presents the ROC-AUC scores for all models evaluated on the Conflict Forecast dataset from 2018 to 2023.



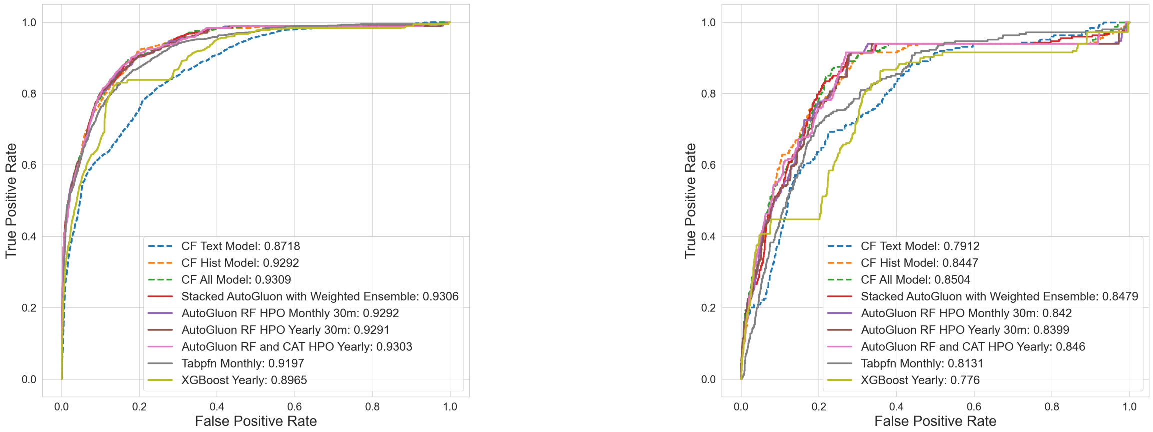
(a) Model performance on Conflict Forecast dataset (2018–2023), all onsets.

(b) Model performance on Conflict Forecast dataset (2018–2023), hard onsets.

Figure B.2: ROC-AUC score comparison of models on the Conflict Forecast dataset for all onsets and hard onsets.

B.3. UCDP Dataset ROC-AUC Scores

This subsection in the Figure B.3 presents the ROC-AUC scores for all models evaluated on the UCDP dataset from 2018 to 2023.



(a) Model performance on UCDP dataset (2018–2023), all onsets.

(b) Model performance on UCDP dataset (2018–2023), hard onsets.

Figure B.3: ROC-AUC score comparison of models on the UCDP dataset for all onsets and hard onsets.