Integrating XAI for Predictive Conflict Analytics

Luca Macis[1], Marco Tagliapietra[1], Alessandro Castelnovo[2], Daniele Regoli[2], Greta Greco[2], Andrea Cosentini[2], Paola Pisano[1], Edoardo Croccetto[1]

[1]University of Turin, [2]Intesa Sanpaolo S.p.A.

{name.surname}@unito.it, {name.surname}@intesasanpaolo.com

In a nutshell

- **Predicting potential conflicts** has played a crucial role in the landscape of peace research since Singer's work in the early 70's.
- The most current and prominent example is the Violence and Impacts Early-Warning System (VIEWS) developed by Uppsala University, `https://viewsforecasting.org/`;
- in collaboration with the Italian Ministry of Foreign Affairs, some of the authors have implemented a **new AI-EWS employing Transformer models**, built upon a multi-headed attention mechanism.[a]
- However, the usage of such sophisticated techniques, if on one side brings the benefit of increased accuracy of conflict predictions, on the other it opens new challenges to be faced. One of such challenges lies in the **inherent complexity of Deep Learning based models**.
- We introduce **XAI approaches** — in particular, those based on **Integrated Gradients** — to enhance the transparency and comprehension of the AI-EWS's outcomes.
- We make use of ACLED dataset[b] to validate our analysis.

[a] Contribution to the VIEWS Prediction Challenge 2023/204. Please see Hegre et al. (Forthcoming) and `https://viewsforecasting.org/research/prediction-challenge-2023`

[b] Armed Conflict Location and Event Data Project (ACLED); `https://acleddata.com/`

Geopolitical data

ACLED dataset:

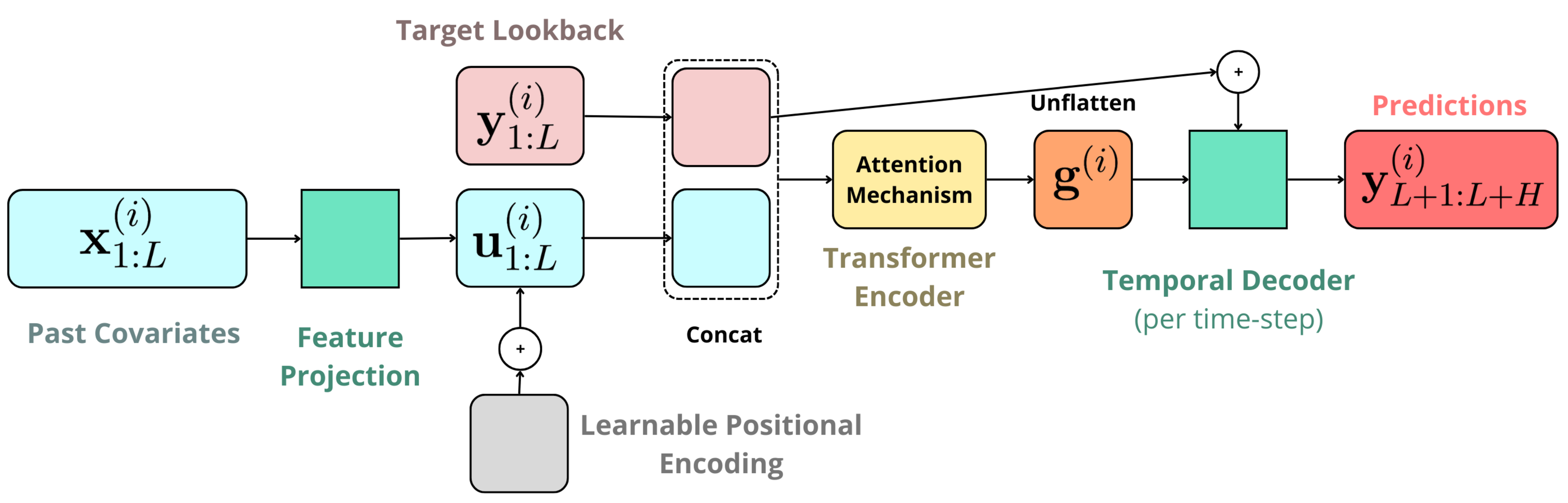
- is regularly updated time-series dataset;
- is publicly accessible via API.
- chronicles various conflict events with distinct descriptions, location, and time, and ensures reliability through a rigorous verification process.
- It's detailed structure and transparency foster academic research and informed decision-making.

We **aggregate data weekly** + **categorise them by event types** $\rightarrow$ 50-valued time-series dataset where each observation corresponds to the **number of event type within that week for that country**.

Country	Date	Armed clash	Air/drone strike	...	*Fatalities*
Afghanistan	2016-12-31	155	13	...	666
Afghanistan	2017-01-07	140	10	...	546

Fatalities represents the target variable, obtained by summing up all fatalities recorded in that week for that country.

Transformer Model



1. Input: $x^{(i)}_{1:L}$ = independent *multivariate* time-series features. $y^{(i)}_{1:L}$ = target variables (# of fatalities).
2. (i) = specific country. $1\ldots L$ = **Lookback** time steps considered. $L = 48$ weeks.
3. Input dimensionality is reduced using a **Residual Block** (i.e. MLP + identity skip connection) $\rightarrow u^{(i)}_{1:L}$.
4. Concatenate $u^{(i)}_{1:L}$ and $y^{(i)}_{1:L}$ and pass them through the **Transformer encoder**.
5. Unflatten + temporal decode the output.
6. Recursively predict $y^{(i)}_{L+1:L+H}$, $H = 12$ weeks horizon.
7. Minimize: NLL Loss$^{(i)} = -\frac{1}{N}\sum_{\gamma=1}^{N}\sum_{j=1}^{H} \log P(y^{(i)}_{\gamma+j} \mid y^{(i)}_{\gamma-L:\gamma}, x_{\gamma-L:\gamma})$.

Transformer hyperparameters

feat_proj	transf_enc_dim	head_size	num_heads	num_transf_enc	dec_dim	dropout
16	16	16	3	1	256	0.1

Integrated Gradients

Objective: attributing the prediction of a deep network to its input features (a.k.a. feature importance).
Axiomatic approach [1]:

1. **Sensitivity**: for inputs and baselines differing in a single feature yet yielding different predictions, the differing feature must receive a non-zero attribution.
2. **Implementation Invariance**: the attributions are always identical for two functionally equivalent network.
3. **Completeness**: attributions add up to the prediction difference wrt the baseline $\rightarrow F(x) - F(x')$.

Contribution of k-th feature to shift model F prediction from baseline value $F(x')$ to $F(x)$:

$$IG_k(x) = (x_k - x'_k)\int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_k}\, d\alpha.$$

Baseline $x' \rightarrow$ column-wise means after rescaling.

Related Works

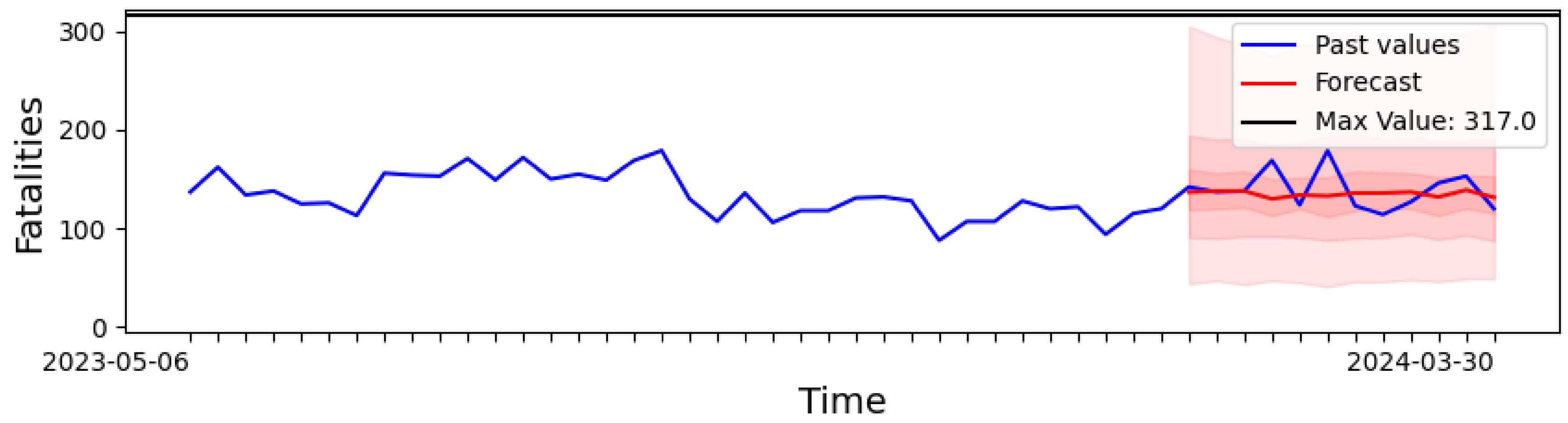
Our architecture has two main components:

1. Time-series Dense Encoder — **TiDE** [2].
2. Use of Transformer for multivariate time-series [3].

In particular, we draw by TiDE the use of **Residual Blocks**, making it straightforward to reproduce simple linear models, but instead of using a Dense encoder, we leverage the potential of Tansformers. Indeed, it has been noted that Transfomers alone cannot effectively model time-series dynamics [4].

Forecast

This study sets out to **forecast the potential number of fatalities from February 13, 2024, to March 30, 2024**[a], focusing on **168 countries** that have recorded **at least one fatality** throughout their historical time series[b].

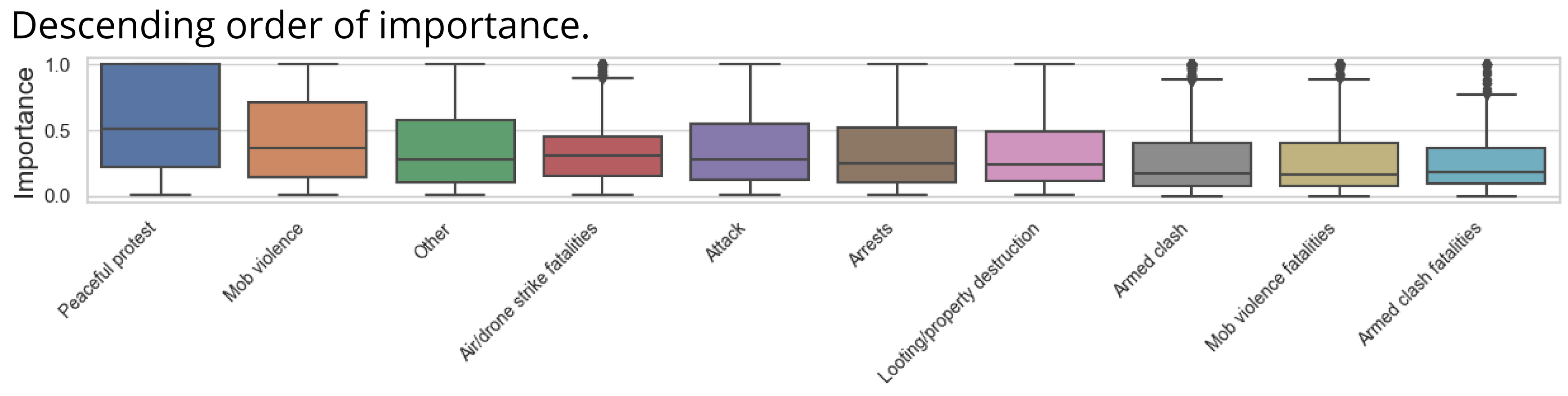


[a] This timeframe spans the most recent twelve weeks of ACLED data available for each country up to April 9, 2024.

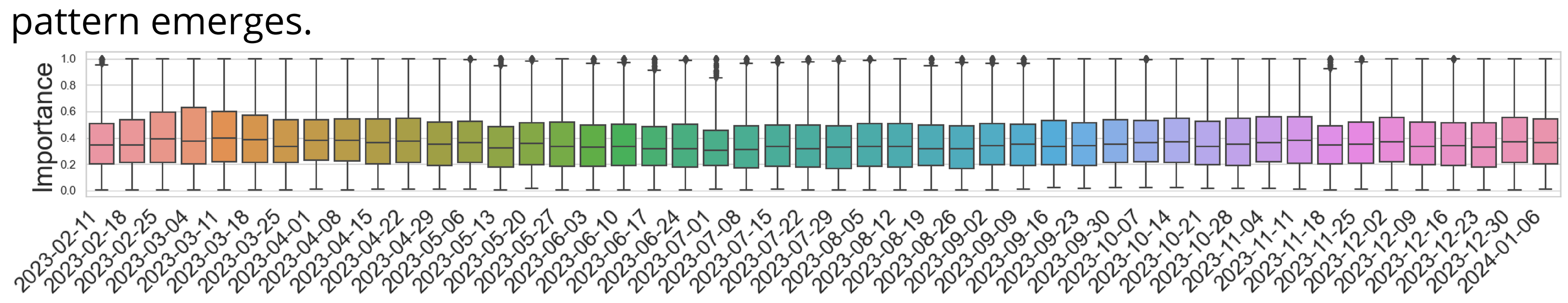
[b] This criterion is crucial as predicting deviations from zero fatalities where no prior occurrences exist is statistically improbable. Therefore, countries are assessed individually based on their historical data.

XAI analysis

- **Aggregated feature importance**. Top-10 Variables for Importance across all countries. Descending order of importance.



- **Temporal Insights** with Integrated Gradients across all countries $\rightarrow$ no clear temporal pattern emerges.



References

[1] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. *Axiomatic attribution for deep networks.* International conference on machine learning. PMLR, 2017.

[2] Das, Abhimanyu, et al. *Long-term forecasting with TiDE: Time-series Dense Encoder.* arXiv preprint arXiv:2304.08424 (2023).

[3] Zerveas, George, et al. *A transformer-based framework for multivariate time series representation learning.* Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021.

[4] Zeng, Ailing, et al. *Are transformers effective for time series forecasting?.* Proceedings of the AAAI conference on artificial intelligence. Vol. 37. No. 9. 2023.