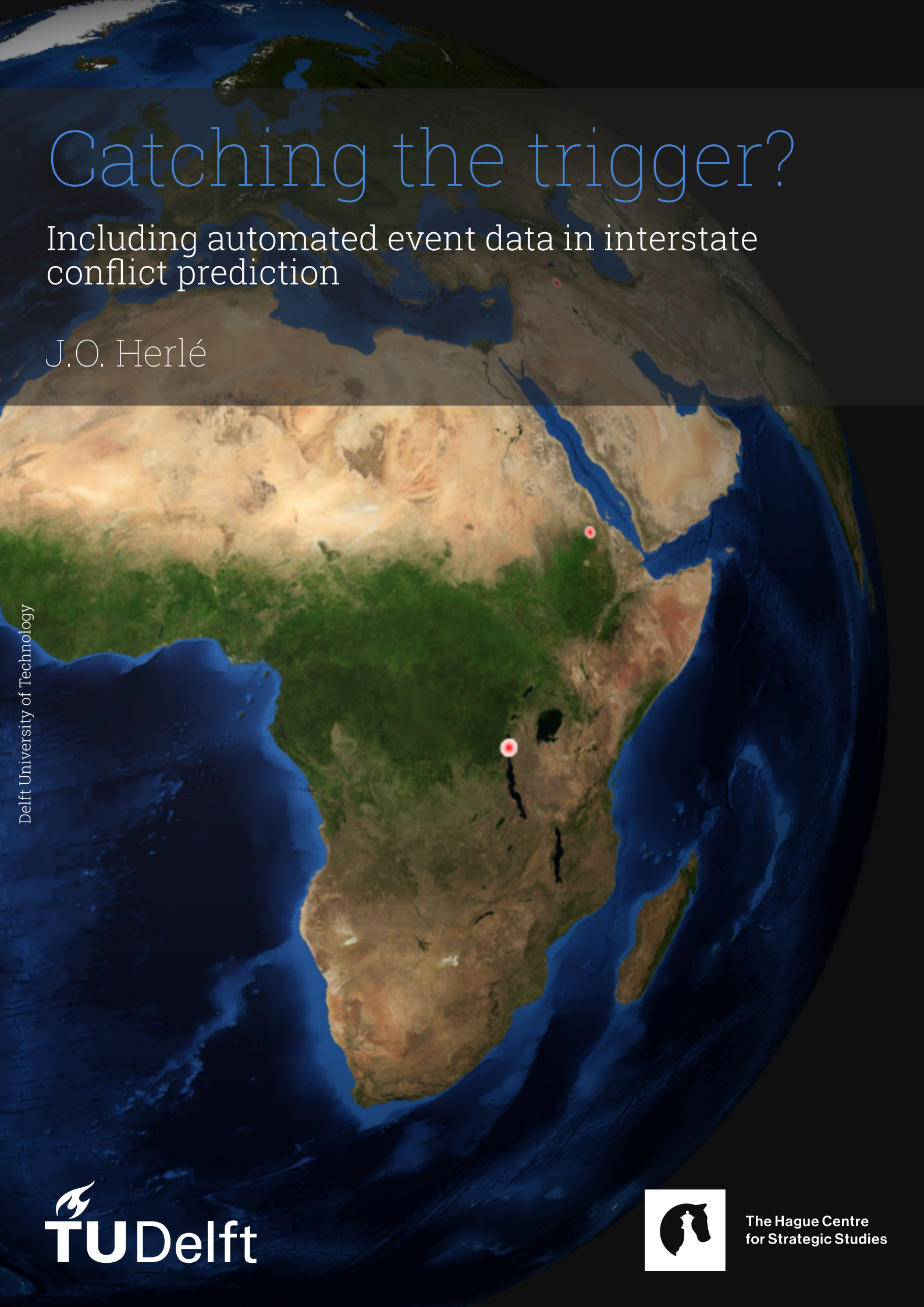
Catching the trigger?

Including automated event data in interstate conflict prediction

J.O. Herlé



Delft University of Technology

TU Delft

The Hague Centre for Strategic Studies

March 2023

The cover image was created using imagery from NASA's Visible Earth collection and was mapped on a globe using maptoglobe.com.

https://visibleearth.nasa.gov/collection/1484/blue-marble?page=1
https://www.maptoglobe.com/#

Catching the trigger? Including automated event data in interstate conflict prediction

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Engineering and Policy Analysis**

Faculty of Technology, Policy and Management

by

Jasper Olivier Herlé

Student number: 4471466

To be defended in public on March 23rd 2023

Graduation committee

Chairperson	: Prof. Dr C.P. van Beers, Section Economics of Technology and Innovation
First Supervisor	: Dr J. Zatarain Salazar, Section Policy Analysis
External Supervisor	: Dr T. Sweijs., The Hague Centre of Strategic Studies

TU Delft

The Hague Centre for Strategic Studies

EXECUTIVE SUMMARY

Violent conflict can be devasting on human life and prosperity. Therefore, both scholars and policy professionals have spent much time and effort in developing early-warning systems. These early-warning systems provide predictions and warnings of conflict. Informed policy makers can subsequently take action. Early or anticipatory action can greatly reduce the costs of responses at a later time (Organisation for Economic Co-operation and Development (OECD), 2009). Therefore, the field of conflict prediction seeks to develop predictive models for diverse types of conflict. These predictive models can subsequently serve as the input for early-warning systems.

The proliferation of machine learning and new datasets pushed the performance of these models forward. Particularly important is the development of automatic event data (AED). These are data coded automatically from news messages and they are thus suited to capture short-term dynamics. These AED complement the traditionally used structural variables, which refer to slow-changing variables. The combination of AED and machine learning has scored several predictive successes, such as the Violence Early-Warning System (ViEWS) project, which predicts state-based conflict in Africa. However, these successes so far have not included interstate conflict prediction. Therefore, the premise of this study is to evaluate the contribution AED makes in interstate conflict prediction. Additionally, this study provides global predictions on a monthly level, whereas interstate conflict prediction before only has used yearly predictions on a global scale. Subsequently, it provides predictions for the occurrence, the onset and the escalation of interstate conflict. Additionally, the project seeks to identify the driving factors in model construction by calculating feature importance scores.

For one part, the model is based on 11 structural features which are causally related to interstate conflict. The other strand of input data consists of 268 features based on the Integrated Crisis Early Warning System (ICEWS) data. These are monthly counts of scraped news events coded according to the Conflict and Mediation Event Observations (CAMEO) coding ontology. Thus, a collection of 279 features predict the outcome variable. The outcome variable itself is based on militarized interstate disputes (MIDs). The models used are random forest and eXtreme Gradient Boosting (XGBoost). The unit of analysis is the dyad month, with data ranging from 1995 to 2014. The peace instances are undersampled to prevent class imbalance since they represent the overwhelming majority of the cases for all three prediction problems. The final modelling dataset consists of 9736 cases, half peace cases and half conflict cases. 830 of these are conflict onset cases, and escalation is measured on a six-point ordinal scale, whereas the other two problems are binary. Finally, the importance of the features during model training is indicated using mean decrease in impurity (MDI) scores, although their bias due to the presence of mixed data types and multicollinearity must be taken into account.

The algorithms are tuned with randomised grid search cross-validation and the trained models are evaluated on a held-out test set. The main evaluation metrics are the F2 score and average precision for the binary problems, and the macro-averaged F1 score for the multiclass problem.

The main results of the study, and their interpretation, are:

1. AED included in the feature space as simple counts of events happening on a specific dyad in a specific month does not improve the predictive power of the structural models across all three problems. This result is contrary to the expectation that AED event data capture short-term dynamics. This may be due to the predictive potential of the data itself, or to the choice of aggregation into monthly counts. Other aggregations or feature selection may improve AED performance.
2. The monthly disaggregated models for conflict occurrence and escalation perform well and are an improvement on the state-of-the-art. Both are, however, relatively uninformative outcome variables, so are not suitable to form the basis of an early-warning system. Nevertheless, a workaround may be viable where occurrence is used to predict conflict onset. Additionally, the escalation models separate different levels of escalation fairly well. They reflect the highest level of escalation achieved in an MID, and can therefore not be used to model intraconflict dynamics,

but the models do indicate that it is achievable to separate MID according to their level of escalation on a monthly level.
3. The predictive power of the onset models is insufficient. Further improvements to the models, such as cost-sensitive training or feature selection may improve the models' performance. The onset models do not lend themselves to early-warning models in any capacity in this state.

The main limitations of this study pertain to the data setup and methods used. First, the model is trained on data only from 1995 to 2014, leaving a large gap between the present years and training data. This means that the model is not suitable for direct real-time prediction, since it has not learned the structures of the last seven years, which is sizeable compared to the twenty training years. Second, the MDI scores are unable to provide robust indications of the most important variables, and third, the model uses an artificially balanced test set, instead of the real-world ratio of peace and conflict cases.

To continue in the field of conflict prediction, the study makes three recommendations.

A. Improve the current models methodologically and implement cost-sensitive training to evaluate whether the lacklustre AED performance can be improved, as well as the onset models' performance.
B. Use Ensemble Bayesian Model Averaging in subsequent modelling efforts for interstate conflict. Ensemble methods, as well as Bayesian methods, have been shown to be effective in similar problems (Hegre, Bell, et al., 2021; Williford & Atkinson, 2019).
C. Host a forecasting competition for interstate conflict onset to acquire a focused and diverse set of efforts. These can then also be used in a model ensemble.

In conclusion, this study tested the predictive contribution of AED in the prediction of the occurrence, onset, and escalation of interstate conflict. Contrary to expectation, AED does not contribute to better performance, at least when aggregated into monthly event counts. On the other hand, the occurrence and escalation models perform well, indicating that monthly predictions are possible for interstate conflict.

Hegre, H., Bell, C., Colaresi, M., Croicu, M., Hoyles, F., Jansen, R., Leis, M. R., Lindqvist-McGowan, A., Randahl, D., Rød, E. G., & Vesco, P. (2021). ViEWS2020: Revising and evaluating the ViEWS political Violence Early-Warning System. *Journal of Peace Research*, *58*(3), 599–611. https://doi.org/10.1177/0022343320962157

Organisation for Economic Co-operation and Development. (2009). *Preventing Violence, War and State Collapse: The Future of Conflict Early Warning and Response*. OECD. https://doi.org/10.1787/9789264059818-en

Williford, G. W., & Atkinson, D. B. (2019). A Bayesian forecasting model of international conflict. *The Journal of Defense Modeling and Simulation*, *17*(3), 235–242. https://doi.org/10.1177/1548512919827659

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to the members of my graduation committee. Tim Sweijs, who has guided me in the world of conflict prediction and has been very involved over the duration of the project. Jazmin Zatarain Salazar, who as my first supervisor allowed me a great deal of freedom in shaping the project, and Cees van Beers, whose suggestions have been greatly valuable along the duration of the project.

At the Hague Centre of Strategic Studies, I want to thank the people of the Data Lab. Maarten Vonk has always been there to explain the Data Lab's policies and systems, and has been a valued partner in discussion. I would also like to extends my thanks to Nino Malekovic for his constructive feedback and stimulating discussion.

Lastly, I'd like to acknowledge Daniella Kranendonk, who has been very friendly, flexible and helpful in coordinating logistics during my internship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

AED	Automatic event data
CoW	Correlates of War
EU	European Union
GDELT	Global Database of Events, Language, and Tone
ICEWS	Integrated Crisis Early Warning System
MID	Militarized Interstate Dispute
PRC AUC	Precision-Recall Curve Area Under the Curve
RF	Random Forest
ROC AUC	Receiver-Operating Characteristic Area Under the Curve
UN	United Nations
VIF	Variance inflation factor
XGBoost	eXtreme Gradient Boosting

***group authors erbij toegevoegd?

1 INTRODUCTION

Early warning systems can predict incoming shocks to policy makers. While these systems have been developed and put in place for phenomena such as tsunamis, droughts, and disease outbreaks (UN Office for the Coordination of Humanitarian Affairs, 2022), effective systems for human conflict are more difficult. But this has not stopped the academic field of conflict prediction nor the policy world from trying to develop models that can predict conflict.

The need for early warning systems is recognised by policy makers worldwide. The international community, particularly the United Nations (UN), has been striving to establish an effective and reliable conflict early warning system for the last 30 years. Starting from Boutros Boutros-Ghali's An Agenda for Peace (UN, 1992) to the present day, UN policy makers and academics have emphasized the importance of having comprehensive early warning systems that can provide policy makers with accurate predictions by gathering and analysing information related to the early warning indicators of conflict. In 2011, Secretary General Ban Ki-Moon and the UN Security Council reconfirmed the need for early warning mechanisms in the Preventative Diplomacy: Delivering Results report (UN, 2011). Several early warning systems have been created, the European Union Conflict Early Warning System being among the few that link warning to action (EU, 2020; Meyer et al., 2019).

The rationale behind early warning systems is that if conflict can be predicted, then informed policy makers can take anticipatory action to reduce the humanitarian impact, adjust policy to better safeguard interests, or in the best-case scenario, even take preventive measures. By early or anticipatory action policy makers can greatly reduce the costs that responses would have taken at a later time (OECD, 2009).

The basis of these systems is often formed by models analysing indicators of conflict. While some models only track the indicators themselves, others can assign probabilities of conflict. Machine learning has greatly improved the performance of these models, but nevertheless, effective predictive early warning systems are scarce for violent conflict. A major reason for this scarcity is the current inability of the field of conflict prediction to predict conflict sufficiently accurately to base policy interventions on.

This research focuses on increasing the reliability and accuracy of the predictive models, contributing in this way to enable effective early action. It does so in the context of interstate conflict, which has received much attention following the invasion by Russian forces of Ukraine. This study tries to predict various aspects of militarized interstate disputes (MIDs)[1] by using a machine learning approach. Once predictive models are accurate and reliable, they can be implemented in early warning systems.

[1] "united historical cases of conflict in which the threat, display or use of military force short of war by one member state is explicitly directed towards the government, official representatives, official forces, property, or territory of another state" (Jones et al., 1996, p. 163)

1.1 Knowledge gap and research questions

Over the last two decades, the literature on conflict studies has seen the growth of conflict prediction from a modest subject to a prominent area of research. This was made possible by the proliferation of large datasets and computational techniques, such as machine learning algorithms. Machine learning now is a very popular approach to conflict prediction. Much of this growth has focussed on intrastate conflict, i.e., insurgencies, civil wars, and other forms of political violence between a government and other actors. Conversely, there is less research on conflict prediction related to interstate conflict. This is mainly due to the relative scarcity of interstate conflict compared to intrastate conflict. This means that the literature on conflict prediction regarding interstate conflict leaves room for improvement. More specifically, no studies to date have tried to introduce automated event data (AED) to capture short-term dynamics that may trigger interstate conflict to erupt. This project tries to fill that gap and additionally strives to increase understanding of which factors are most influential in the occurrence, onset, and escalation of interstate conflict. To that end, this project tries to answer the question:

"How can machine learning techniques and automated event data be employed to better predict and understand the onset and escalation of militarized interstate disputes?", which can be divided into the following sub-questions:

SQ 1. "What factors are likely to be associated with MID onset and escalation?"

SQ 2. "What algorithms and methods are best suited to predict on combined structural and automated event data?"

SQ 3. "Can predictions for MID onset and escalation be improved with automated event data?"

SQ 4. "What factors are important in predicting MID onset and escalation?

The objectives of answering these questions are 1) to improve on existing models in the extant literature, 2) to refine the understanding of conflict onset and escalation, and finally, 3) to provide the basis of an early warning tool for policy makers for the onset and escalation of interstate conflict.

The terms prediction and forecasting have been used interchangeably in the field of conflict studies. Prediction is defined as the assignment of an outcome probability to realised or unrealised events, including model estimates, but also prediction based on expert knowledge. Forecasting is a term that is also often used and is here taken to mean predictions about unrealised outcomes given model estimates from realised data, following Hegre et al. (2017). Seen this way, prediction is the first step in the chain, forecasting takes tested predictive models to the uncertain future, and early-warning systems do this in a systematic way.

This study will focus on conflict *prediction*, and provide the basis for later iterative efforts towards forecasting and early-warning systems.

1.2 Research approach

The investigation of the research question is empirical and follows a data-driven approach, with some qualitative elements.

To find out what factors likely are associated with MIDs, the literature on the causes of conflict and the factors correlated with it has been reviewed. Then, a quantitative approach, machine learning, is used to find the relations with which to predict interstate conflict and to build the models to predict the occurrence, onset, and escalation of conflict.

1.3 Link to EPA programme

The core of the Engineering and Policy Analysis (EPA) programme is the aim to provide policy advice in complex socio-technical systems, through a variety of methods: qualitative, but predominantly with a quantitative focus. The objectives of this research fit the programme well, since it aims to provide information about conflict onset and escalation to public policy makers, as well as contribute to the field of conflict studies. Although the system itself is

mainly a social system, the data analytical methodology and policy component anchor the research within the programme.

Additionally, this project takes both a systems and a multi-actor view towards the problem. It looks for regularities on a system level, and it analyses dyadic actor behaviour in order to construct a dyadic dataset. Finally, this research contributes to the solution of the Grand Challenges, which are central to the EPA programme. More specifically, it ties into Social Development Goal 16: the promotion of peaceful and inclusive societies (United Nations, 2022), since it is explicitly aimed at increasing the anticipatory capacity, and as such, the capacity for early action of public policy makers to interstate conflict.

1.4 Report structure

This thesis report is structured into five chapters. Chapter 1 details the context of the problem, and explains the setup of the study. It describes the knowledge gap and research questions, as well as the objectives of the study. It also outlines the research approach it introduces the report's structure. Chapter 2 provides an overview of the scientific literature on conflict prediction. It gives a historical overview, discusses major debates, and identifies the knowledge gap in the literature. Chapter 3 presents the methods used in the study. It describes the data used, what operations have been performed on the data, and what their structures are. It then outlines which algorithms are used, and how they are used. Finally, it then portrays the way that the models created have been evaluated. Chapter 4 presents the results of the predictive models, and analyses their performance. Here, the question of whether AED contributes to modelling is evaluated, as well as what factors are most influential in prediction. Subsequently, Chapter 5 then discusses the findings and their implications. It also mentions the limitations of the research and provides recommendations for further research. Finally, the conclusions are presented in Chapter 6.

2 LITERATURE REVIEW

A fundamental goal of peace research is to predict peace and conflict (D. Singer, 1973). This chapter provides a review of the field of conflict prediction.

The chapter is organised as follows. First, a historical overview highlights the most consequential developments in the field. Second, the role of prediction in the field is discussed. Third, the positions and challenges around the fundamental predictability of conflict are discussed, before concluding by sculpting out the contributions of this study within the niche of interstate conflict prediction.

This review aims to provide an overview of the field of conflict prediction. Additionally, there is a large literature on the covariates and causal theories of the various types of conflict, but these are not in the scope of this review.

2.1 Conflict prediction: a historical perspective

There are several ways to treat the history of conflict prediction. This review uses the three generations in the field as identified by Hegre et al. (2017).

The first generation of conflict prediction was inspired by the works of Wright (1942/1965) and Richardson (1949/1960), who wrote sizeable studies dedicated to the systematic and statistical analysis of war. With his statistical analysis of the phenomenon of war, and the international relations field in general, Wright confirmed earlier ideas about conflict which had hitherto only been treated theoretically. Examples are that constitutional government is more favourable to peace than absolutism, or that rules of war and of neutrality have no significant effect in limiting war. Richardson, comparably, treated historical conflict in a mathematical manner. He found amongst others that the relationship between the frequency and size of conflict, measured in the number of casualties on a logarithmic scale, follows a Poisson distribution, with the smaller conflicts occurring much more frequently than the larger ones. However, rather than with their exact findings, the works of Richardson and Wright hold significance for the field of conflict prediction by being the first major statistical works on conflict, opening the door for others to predict conflict based on empirical regularities.

Following this systematic approach, one of the earliest works that included a predictive component was a study of the Vietnam war (Milstein, 1974). Using detailed data on troop numbers, abducted civilians, number of bombing sorties, and more, Milstein used the outcomes of his statistical models in a simulation model with the aim of predicting escalation in the war, depending on the policy choices that were made. The ultimate goal was to evaluate what would have happened, would US policy have been more hawkish, or dovish during the war[2]. Simultaneously, Choucri (1974, p. 80) started promoting forecasts of international relations, calling for systematic evaluations of the

[2] The term hawkish refers to hardliners in foreign policy matters, while dovish refers to policy makers that are more inclined towards cooperation.

effects of current trends and different policy options, naming this the "forecaster's contribution to planning". This task was taken up amongst others by Azar et al. (1977), who predicted crises in the Middle East based on structural data, but also on an early form of event data, finding that accurate predictions of crises could in principle be possible. Later, Azar set up the COPDAB database as a template for curated and fine-grained event data (Azar, 1980). Another important project was the Correlates for War (CoW) project (Small & Singer, 1982), which aimed to systematically collect data about variables theorised to be connected to conflict. One of the aims of the project was to enable better models for conflict prediction (J. D. Singer & Wallace, 1979), and it became an important, if not the most important source for conflict data in the ensuing decades.

Thus, the first generation of conflict prediction was inspired by the first major statistical works on conflict. It saw the identification of its possible usefulness, and the start of data-collecting efforts aimed at facilitating this process. Nevertheless, after these initial efforts, the interest in conflict prediction faded[3]. The years throughout the late 1970s and early 80s reflected a pause in the field, until the next generation of conflict prediction gained popularity in the 1980s.

The second generation of conflict prediction was characterised by two main developments. These were the introduction of game theory, and the use of econometric models, which went hand in hand with the further development of event data. The introduction of game theory took place slightly earlier than the econometric strand of work, and is presented first.

Game-theoretic models were pioneered by Bruce Bueno de Mesquita (1980, 1985). By eliciting information from expert knowledge about system structure, actor preferences and their power, he was able to construct a simulation model and make predictions with it. With these models, different aspects and theories of actor behaviour could be tested in a simulated reality, and their predictions could be evaluated. Bueno de Mesquita's models in the 1980s focussed on the utility that actors expected to gain in a conflict, but later models included other aspects such as coalition-forming or spatial effects in the context of European Union (EU) voting dynamics (Achen, 2006).

Second, new econometric approaches were introduced starting in the 90s, and being applied to many different aspects of conflict prediction. The goal of these approaches was to make use of new methods and data available and find relations between predictors and conflict data with the aim of extrapolating these relations into the future, as well as analysing the relations by themselves. Before this time, many studies took the form of relatively simple approaches, such as logistic regression, which remained popular during this period. Oneal and Russet (1997) used pooled logistical regressions to test the notion of the liberal peace, finding support for the theory[4]. Another example is Hegre (2008), who demonstrates with a logistical regression that more powerful states experience more conflict. However, there were also more sophisticated econometric methods employed. A starting point was provided by Schrodt, who first used classification trees and then neural networks to predict the outcomes of interstate conflict (Schrodt, 1990, 1991). He found that both the trees and neural network performed better than the oft-used discriminant analysis, but were equal to multinomial logit models. Subsequently, Beck et al. (2000) also used neural networks to assess the effect of commonly theorised predictors of conflict where the ex ante probability of conflict is high, showing the effectiveness of the method for the task. A few years later, Marwala and Lagazio (2004), used Bayesian neural networks to predict the onset of militarized interstate disputes using four structural variables to show the applicability of these methods to MID onset prediction. Somewhat later, they present a volume on the various other computational techniques to model interstate conflict and how they may be used for early-warning systems (Marwala & Lagazio, 2011). They find that these computational or econometric methods are a powerful approach and they present support for a number of variables to influence interstate conflict, such as distance, relative power and common alliances. Finally, Brandt and Freeman (2006) present further support for the use of Bayesian methods by construing policy counterfactuals for the Israeli-Palestinian conflict using Bayesian time-series modelling. The use of these methods[5] constituted an improvement over the first generation, but the second generation was characterised by two more, interlinked, developments.

[3] For a further review of forecasting studies of the first generation, see Choucri and Robinson ((1978), who wrote a book containing a survey of the field, and also Singer & Wallace (1979) for a review of studies pertaining to early warning indicators and forecasting methodology.

[4] The term liberal peace is derived from classical liberal 19th-century political thought and refers to the absence of conflict in economically interdependent states. See Schneider et al. (Schneider et al., 2003) for a review of the concept.

[5] In the wider field of conflict studies, other quantitative studies without a focus on forecasting were also common. Two famous examples are the civil war models by Laitin and Fearon (2003) and Collier and Hoeffler (2004). However, these were sometimes used to form predictions still (Ward et al., 2010).

The first of these was the initial development of automatically coded event data, responding to a need for more and more fine-grained data in addition to the coarse aggregations of structural variables. In the early 90s, Philip Schrodt and his team published the first automated event coding scheme, named the Kansas Event Data System (KEDS) (Schrodt et al., 1994; Gerner et al., 1994). Previously, graduate students created event data by manually filling in coding schemes, but KEDS automatically transformed Reuters news headlines into data points. Now, for the first time, automated coding could drastically increase the volume and coverage of the event data. Later projects in the 2000s would build on this innovation, by expanding in languages and news outlets covered, as well as increasing the volume of the data (Schrodt & Yonamine, 2016).[6] Finally, even though the development of the coding schemes above was a welcome start, the demand for spatio-temporally disaggregated data was still increasing (Cederman & Gleditsch, 2009; Weidmann & Ward, 2010), even more so since Schrodt and Gerner (2000) had demonstrated that disaggregated event data could help form early warning indicators.

In conclusion, the second generation of conflict prediction had delivered a new game-theoretic approach, more sophisticated methods, and the initial stages of automated event data. And although the field was able to identify conflict risk, the exact onset of conflict remained a challenge.

The third generation of conflict prediction, starting in the late 2000s, is characterised by three developments. First, the policy community took further interest in the field, second, the development and use of spatio-temporally disaggregated data advanced significantly, and third, prediction entered mainstream political science.

The promise of conflict prediction to forecast instances of political violence resulted in an increased interest from the policy community in the 2000s. The first major effort was the Political Instability Task Force, which was a CIA-sponsored attempt to globally predict political instability at the country level with a two-year lead time. The project was novel in its scope and ambition, eventually achieving 80% accuracy on political violence onsets over the period 1955 to 2003 (Goldstone et al., 2010). Another notable collaboration between government and academia was the Integrated Crisis Early Warning System (ICEWS) project, which also presented an improvement in the use of disaggregated event data (O'Brien, 2010). Sponsored by the Defense Advanced Research Project Agency (DARPA), the project combined two modelling efforts. In the first place, ICEWS made extensive use of natural language processing to extract events from news sources worldwide, but more importantly, a suite of predictive models was built based on this and other data streams. The predictions of these models were combined via Ensemble Bayesian Model Averaging, allowing different modelling efforts to combine their insights (Montgomery et al., 2012). Thus, the ICEWS project presented a step forward in using AED and model ensembles. Other studies moving away from the yearly format were Ward et al. (2013), who predicted political violence onsets on a monthly basis, as well as Scharpf et al. (2014), who predicted daily risk estimations of massacres in the Syrian civil war.

In the ensuing years, leading up to the present, predictive efforts have proliferated. Various efforts have focussed on early-warning indicators (Blair et al., 2017)), civil war (Chiba & Gleditsch, 2017; Daxecker & Prins, 2017), and geographically disaggregated forecasts (Witmer et al., 2017). Especially notable is the iterative effort of the ViEWS project ((Hegre, Bell, et al., 2021; Hegre et al., 2019; Hegre, Nygård, et al., 2021)), focussing on conflict in Africa, and using different constituent models and different units of analysis to accurately predict the onset of state-based conflict, as well as its location, setting the benchmark for other studies (Brandt et al., 2022).

The acceptance of prediction into mainstream peace research that characterises the third generation of conflict prediction was greatly aided by Ward, Greenhill & Bakke (2010), who made the case that prediction is a better evaluation of theory than the then-usual description of statistical significance (Schneider et al., 2011). This debate is further explored in the next section.

[6] Subsequent projects are the Protocol for Nonviolent Direct Action (PANDA), and the Integrated Data for Events Analysis (IDEA) scheme built on top of it (Bond et al., 2003). Additionally, the Textual Analysis by Augmented Replacement Instructions (TABARI) coder, which is based on the KEDS scheme, allows coding from languages other than English by automatic translation (Schrodt, 2001). Finally, the Conflict and Mediation Event Observations (CAMEO) coding scheme (Gerner et al., 2002) contains a very elaborate event ontology and is the codebook for the ICEWS project data used in this study.

2.2 What role should prediction play in conflict research?

The role of prediction today is twofold (Hegre et al., 2017). The first is to provide timely warnings about events taking place in the world, so that policy makers may act on it (Harff, 2003). The second is as a tool of theory evaluation.

However, evaluating predictive power has become a mainstream test for theoretical models only comparatively recently. Traditionally, theory evaluation had been the domain of statistical significance, p-values, and of case-studies. Although basing confidence in a model's explanatory power on the statistical significance of its variables runs the risk of identifying significant, but small and sometimes irrelevant causal relations, this has been an important test of theoretical validity in the field of conflict studies for decades. Ward (2010) made the case that this approach was inferior to evaluating predictive power by demonstrating that two famous models of civil war did not predict better than a single-variable baseline model when their two most important predictors were deleted from the model, even though many statistically significant variables remained. Thus, prediction provides a valuable addition to significance hypothesis testing for testing theories rigorously against real-world behaviour (Ward, 2016). It should be noted however, that prediction should not be the only evaluator of theory. Some processes are highly path-dependent and can typically only be explained ex post, such as evolutionary processes in nature (Cederman and Weidmann, 2017).

2.3 Prediction in interstate conflict

Up until this point, this review has treated the field of conflict prediction as pertaining to the same predicted outcome. However, this specific study aims to make a contribution in the prediction of interstate conflict. A separation can be made between intrastate, and interstate conflict. The former includes civil wars, coups, violent government crackdowns, genocides and more, whereas the latter refers to violence between the representatives of two nation states. Of the two, intrastate conflict has received more attention. The reason for this is that there are more data points available in the historical record for instances of intrastate violence. Coups, civil wars, and politicides are more frequent than interstate violent conflict. The ViEWS project, for example, makes predictions of intrastate conflict, bringing interstate conflict under the category of state-based conflict (Hegre, Bell, et al., 2021), due to their use of the UCDP data which does not collect data on interstate conflict specifically (Sundberg & Melander, 2013). Thus, interstate conflict prediction is an under researched area where implementing the notions of progress in intrastate conflict prediction could lead to better prediction outcomes.

Within the prediction of interstate conflict, two questions have been long-standing research aims. The first relates to the predictability of the onset of a conflict between two states, and the second relates to the exact timing of escalation within such a conflict. Additionally, both in intrastate and interstate conflict prediction, researchers have struggled with the problem of including temporally and spatially disaggregated dynamics into their models. The use of structural variables, which change only slowly, has been the main approach to prediction, but these variables do not allow disaggregated dynamics to be included. Various types of data have been used to resolve this information problem, among which AED and event data in general. While intrastate conflict prediction has made significant progress in using disaggregated data[7], less headway has been made in interstate conflict prediction regarding this topic. The following paragraphs will detail the state-of-the art of interstate conflict prediction, with a focus on the inclusion of disaggregated data.

Studies into conflict onset were carried out using various types of data inputs. Some used machine learning techniques on structural variables (Beck et al., 2000; K. S. Gleditsch & Ward, 2013; Williford & Atkinson, 2019), or on financial data (Chadefaux, 2017a). Chadefaux (2014) used a form of AED in the shape of news counts to predict interstate war onsets with reasonable accuracy. The best performance for global models of conflict onset, at the yearly level, is represented by Williford and Atkinson (2019), who use a Bayesian forecasting model to achieve an accuracy of 80 per cent at predicted onsets, although at the cost of many falsely predicted conflict onsets. Progress in this matter has been slow however, due to the small number of conflict onsets compared to the instances of peace. Other studies have thus focussed on the concept of conflict occurrence first, as its instances are more numerous than conflict onsets. Two attempts by Marwala and Lagazio (2004, 2011) using various computational techniques, such

[7] For intrastate conflict prediction using disaggregated data, see Hegre et al. (2021), Chadefaux (2022), Brandt et al. (2022), and Vesco et al. (2022) as notable examples.

as Bayesian neural networks and genetic algorithms. They were followed by Stodola et al. (2021), who improved their scores for conflict occurrence using the random forest algorithm.

Finally, the prediction of escalation within a conflict presents another problem. A main difficulty is the scarcity of data on intra-conflict dynamics on a global scale, which is necessary to this end. Additionally, Braitwaite and Lemke (2011) noted that there is no consensus concerning a measure of conflict escalation, as well as few robust relations across different measures. Henrickson (2020) did construct a prediction of the number of battle-deaths intended to further research into the bargaining theory of war, but which may be used as an ex ante expectation of interstate escalation when forecasted. Nevertheless, there are no studies predicting interstate conflict escalation on a global scale to this date.

In conclusion, no study has been undertaken to combine AED with structural data to predict all problems of interstate conflict occurrence, onset, and escalation. Chadefaux (2014) included interstate wars, but no interstate conflicts below the 1000 battle-deaths threshold. Thus, the usefulness of AED in interstate conflict prediction requires further research. This study will contribute to that knowledge gap by testing whether including AED in structural models for interstate conflict occurrence, onset, and escalation can increase the models' performance. Finally, the second contribution this study makes is assessing the possibility of predicting interstate conflict at the monthly level, as current efforts are focussed on the yearly level.

2.4 Can conflict be predicted?

Although conflict prediction in general has made significant steps forward, the question whether conflict can be predicted remains an open question. This section will present a rationale for the endeavour of conflict prediction, as well as the fundamental issues in conflict forecasting. It concludes by providing an overview of characterisations of the nature of conflict.

Whether conflict is inherently predictable or not, has sparked much debate. As illustrated by Tetlock (2005/2017), experts in political science have a poor track record of predicting political events. For some, this indicated that at its very core "research aimed at political prediction is doomed to fail. At least if the idea is to predict more accurately than a dart-throwing chimp" (Stephens, 2012, para. 10). Responding to Stephens however, Hegre et al. (2017) pointed to improvements in data, methods and theory that may eventually allow policy relevant predictions. This lack of past predictive performance is not a call for abandoning prediction for Ward, Greenhill & Bakke (2010), firm proponents of prediction methods. They build a case for a more formalised prediction procedure for theory evaluation, noting a lack of predictive power of current models. Even though doubts persist on the predictability of conflict, it is nevertheless worthy of pursuit, as Jay Ulfelder (2014, para. 4) put it into words: " I land on a pragmatic rationale for emphasizing prediction as a means of validation: it works better than the alternatives".

Nevertheless, four fundamental issues make conflict prediction particularly challenging. First, the structure of the international system is not set in stone, but is constantly evolving. This has two implications. The first is more pragmatic, being that a changing structure changes the setups of most predictive studies, who are based on a fixed geopolitical structure (Cederman & Weidmann, 2017). For example, no model created and trained during the existence of the Soviet Union and treating it as a single country could have predicted a war between Russia and Ukraine, simply because it could not conceive of those entities. Second, beyond changes to the international state system, state behaviours themselves may change drastically, rendering relations learned before that shift less useful (Chadefaux, 2017b). For example, the creation of atomic weapons fundamentally altered the calculus to go to war with an equally armed opponent. Changes such as these alter a models outcomes drastically. The second fundamental issue is data quality. Timing and location of conflict are prone to measurement error, as are the variables used to predict conflict (Cederman & Weidmann, 2017; Chadefaux, 2017b). Third, as Cederman and Weidmann (2017) note, conflict is inherently complex and is an environment with strategic actors. Actors change their behaviour based on their perception of the future, meaning that forecasting models could disturb the relations they themselves have found when their insights are put to use. Chadefaux (2017) adds to this that states employ mixed strategies in their behaviour to prevent predictability of their actions, thus purposefully distorting any relations found in historical data.

Next to the challenges in conflict prediction, the very nature of conflict itself is not clear. Conflicts may ultimately be predictable, regular occurrences, such as the behaviour displayed by clocks. Otherwise, they may be disorderly phenomena, irregular and "more or less unpredictable"[8], while we can still learn about the underlying probability distributions (Chadefaux, 2017). Another possibility is that it not only is conflict unpredictable, but that the factors on which its probability would depend also are unknown. In this case, conflict would be a black swan (Taleb, 2007).

[8] For the concepts of clocks and clouds referred to, see Popper (1979).

Chadefaux (2017) outlines several approaches to finding out that question, mainly by carefully examining past and current predictive performance, where a plateauing discipline might point to fundamental limits. Some nuance is provided by Gleditsch (2017), who warns of excessive focus on unpredictable black swans, and focussing on more regular white swans[9], providing the valuable notion that not all conflicts are equal in nature. As such, many studies continue predictive efforts and keeping track of progress can allow us to discern whether conflict ultimately is a predictable phenomenon.

[9] In his study, he refers to white swans as having "many regularities in their behaviour or whereabouts at specific time points" (K. S. Gleditsch, 2017, p. 1).

3 METHODOLOGY

The methodology chapter presents the details of the methods used. It first lists the types of data and how they have been used in the study in the Data section. Appendix A. Data is an addition to this section. Subsequently, the modelling process is introduced in the Modelling and Prediction section, going into detail of the logic behind the model selection, hyperparameter tuning and model evaluation.

3.1 Data

3.1.1 Data structure

The research design is centred around the non-directed dyad month as the unit of analysis. A dyad is a pair of countries. To account for various different points in time, a dyad can be transformed into a dyad month, which refers to a specific month on a specific dyad. An example would be France – Belgium, March 1998. Non-directed then refers to the fact that there is no directionality in the unit of analysis. Directed dyads would be France → Belgium, and conversely, Belgium → France. Instead, a non-directed dyad solely refers to a combination France – Belgium. This format aggregates directional aspects, such as who-did-what, into dyadic aspects. The main reason for choosing undirected above directed dyads, is that disaggregating dyadic data into directed dyads at a monthly level would make the data even more sparse than it would be at an undirected level. This paucity of data would make prediction in an already data-poor environment even more difficult.

The dyads, or country pairs, form a new data point for every month. This is done on a global scale, for 187 states, in the period 1995 – 2014. The time period is mandated by the time limits of the available data. The result is a raw dataset structure of 4,173,840 dyad months.

The input data for prediction come in two flavours. Eleven features are made up by operationalising structural variables, variables that do not change a lot in a short time span, and that set the conditions for potential conflict. The second data stream consists of automated event data, which automatically codes news events from all around the world. The idea under review is that conflicts are in large part caused by dynamic triggers, which cannot easily be measured with structural variables. The main question in this study is whether automatic event data can catch these dynamic processes and improve upon existing models of interstate conflict prediction.

3.1.1.1 On using all possible dyads

The final dataset contains all possible dyad months in the period 1995-2014, their structural features, and the event counts of the events happening on those dyads. The total number of rows is larger than four million, the overwhelming majority of which are populated by peaceful dyads. Additionally, these dyads may not even interact much, thus making them unlikely to experience conflict in the first place. Some studies choose to model using only

a subset of 'politically relevant dyads' (Stodola et al., 2021), as to increase the ratio of relevant to irrelevant cases, or decrease the need for data collection. The latter reason is not applicable to this study, which uses available global data, but the former is.

Nonetheless, using only politically relevant dyads has three drawbacks. First, any conflict episodes outside of the politically relevant dyads will not be included in the training set. Subsequently, the model will find it difficult to predict those cases, given that may not have seen a similar conflict in its training data. Second, including only politically relevant dyads invites selection bias into the results regarding which features are the most influential predictors. Although this has been found not to constitute a major problem for well-researched correlates of conflict (Lemke & Reed, 2001), if it can be avoided, then that would be an advantage. And finally, even with the politically relevant dyads, further measures would have to be taken to deal with class imbalance within the targets' classes.

3.1.2 Sorts of data

3.1.2.1 Target variables

The variables to be predicted are commonly referred to in the machine learning field as targets, while the predicting variables are known as features. This mirrors the statistical terminology of dependent and independent variables, or the terms input and output variables associated with models. In essence, all refer to the same concept of input variables that are able to say something about the variable of interest.

In this project, the main interest lies with predicting the 1) onset and 2) escalation of interstate conflict. These are operationalised in two separate target variables. Additionally, 3) the occurrence of conflict is included as another target variable. This is to better ground the model in existing literature, where occurrence, not onset, is traditionally the variable of interest. This traditional focus, however, is not due to lack of theorical or empirical interest in the onset of conflict, but rather to the larger volume of historical conflict occurrence cases. After all, a single conflict onset can embody many yearly or monthly conflict occurrence cases.

These three target variables then need to be sourced from a dataset on historical conflict. There are a multitude of conflict datasets available, each with their own focus and coding decisions. Consequently, choice for a data source is project specific. The conditions and criteria for this choice are as follows.

i. Must explicitly code for interstate conflict
ii. Must allow for disaggregation at the level of the dyad month
iii. Must offer a way to construct a measure of escalation
iv. Contain as many conflict episodes as possible
v. Must be global and extend for as much time as possible after 1995

In the end, the MID Dyadic Dataset best meets all demands, and thus was the dataset chosen. It runs from 1816 to 2014 and captures MIDs. For an overview of the datasets considered, see Appendix A. Data.

The target variables are then operationalised as follows. Occurrence is coded 1 when two states are involved in an MID in any given month. Onset is also coded on a binary level, with the starting month of a new MID being the positive case. The operationalisation of the escalation target variable deserves more detail, which is examined in the section below.

To get a better intuition about the target variable, occurrence of conflict in this case, Figure 3-1 presents the number of dyads that experienced an MID in any given year. It is interesting to see large fluctuations, even outside of both World Wars. Note however, that this study does not account for the temporal structure between the units of analysis, the dyad months. The effect of time is accounted for, however, in the feature Dyadic peace years (see the following section), and by lagging the event features by a month.

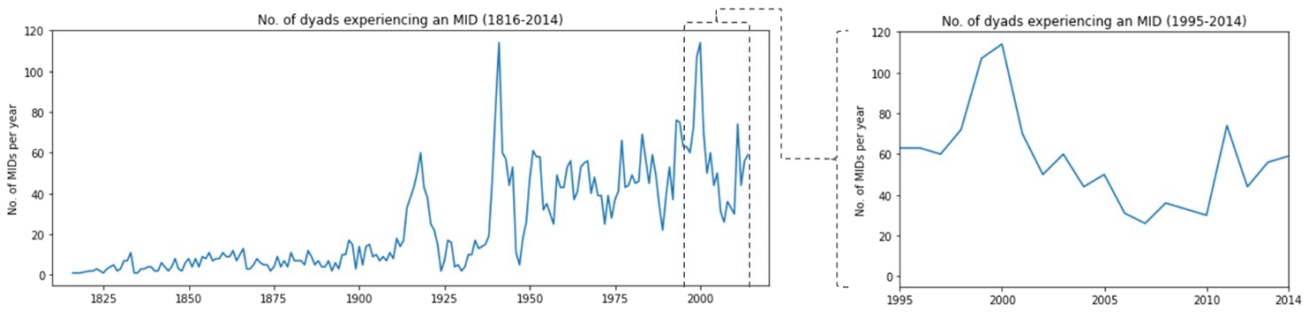


Figure 3-1. MIDs through time

Operationalising escalation

There is no agreed upon measure for escalation in the conflict studies literature (Braithwaite & Lemke, 2011; Terechshenko, 2020), although there are recurring elements, such as thresholds for numbers of fatalities, the use of force, and reciprocal or unilateral actions. Apart from hampering progress in research on escalation processes due to incomparability, this also means that this project must make its own decision regarding an escalation measure.

The most-used measure in the field is the hostility variable by the CoW Project (Jones et al., 1996). It recognizes five ordinal categories. The absence of militarised action, the threat to use force, the display of force, the actual use of force, and interstate war, where participants must suffer at least 100 battle-deaths, and deploy at least 1000 troops in battle. This ordinal scale however, is criticised due to a lack of scientific evidence. For example, is a show of force more hostile than a threat to declare war (Diehl & Goertz, 2000; Maoz, 1982)?

We compare the CoW measure with two other measures that tweak this operationalisation. First, Braithwaite & Lemke (2011) test commonly used predicting variables with several different measures of escalation. They look at escalation first by seeing whether any force was used, then by whether any actions taken were reciprocated, or finally by noting the level of fatalities (0, 250, 1000). Second, Bennet and Stam (2004) collapse the divide between threats and displays, and distinguish between one-sided, or reciprocated use of force. An overview of the options considered is presented in Table 3-1.

For the purposes of machine learning, it is important that the various classes contain enough cases, preferably evenly spread. Additionally, the more granular the measure, the more precise predictions can be made. Based on these considerations, an alternative measure is used, that distinguishes between threats, displays, one-sided and reciprocated use of force, but not between different fatality levels. Thus, escalation is coded on an ordinal scale, where 0 means no militarised action, 1 a threat to use force, 2 a display of force, 3 a one-sided use of force, 4 a reciprocated use of force, 5 interstate war. Note that this ordinality is a rough measure, which is not to be used for assessing theoretical claims on escalation processes, but which is useful for our predictive purposes.

Table 3-1 Options for operationalisation escalation
(undirected dyadic MIDs, relevant states only, 1995-2014)

A. CoW		*B. Braithwaite & Lemke*		*C. Bennet & Stam*		*D. Operationalisation used*	
1 No MID	many	No MID	many	1 No MID	many	1 No MID	many
2 Threat	43	Use of force	610	2 No UoF	486	2 Threat	43
3 Display	443	Reciprocated	323	3 One-sided UoF	333	3 Display	443
4 Use of force	610	Fatality level:		4 Reciprocated UoF	277	3 One-sided UoF	333
5 War	46	0	960	5 War	46	4 Reciprocated UoF	277
		0-249	89			5 War	26
		250-999	8				
		1000+	28				
Total	1142				1142		1142

The escalation operationalisation has an important consequence for the precision with which the models can predict escalation in time. In the CoW data, escalation is measured at the dispute level. The variable hihost indicates the highest level of hostility experienced during the entire dispute. As such, the algorithms can only predict the highest level of hostility experienced during a dispute, not the course of hostility levels throughout it. A hypothetical conflict illustrates the concept in Figure 3-2.

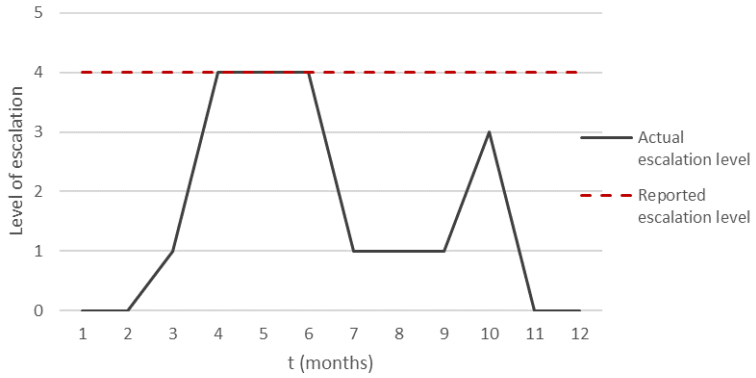


Figure 3-2. Escalation levels during an eight-month conflict

When transformed into a dyad month format, the target variables' class distributions become rather imbalanced, as can be seen in Table 3-2. This dataset is then undersampled with a 1:1 Occurrence – No occurrence ratio, which is detailed in subsection 3.1.3.

Table 3-2. Distribution of classes, raw data

Occurrence		*Onset*		*Escalation*	
0	4168944	0	4173007	0 (no MID)	4168944
1	4896	1	833	1 (threat to use force)	153
				2 (display of force)	1263
				3 (one-sided use of force)	1292
				4 (reciprocated use of force)	1824
				5 (interstate war)	364
Total	4173840		4173840		4173840

3.1.2.2 Structural variables

Structural variables refer to slow-changing variables that shape the conditions in which conflict may or may not erupt. Traditionally, it has been a focus of academics in conflict research to disentangle the processes leading to conflict and conflict escalation by carefully examining any statistically significant effects these structural variables have on the variables of interest. As such, they also formed the starting point for predictive studies, since a wealth of theoretical links and carefully crafted datasets were already available.

In this study, eleven input variables are included, as presented in Table 3-3. They were selected based on theoretical links with the processes of conflict onset and escalation, and all take the shape of the dyad month. For a detailed background on arguments for inclusion and specific operationalisation, see Appendix A.

Table 3-3 Overview structural variables

	Variable	*Operationalisation*	*Data source*
Political			
	Existing rivalries	1 if two states see each other as a significant threat, else 0	Thompson rivalry dataset, v2022
	Joint democracy	1 if two states score 6 or higher on the 0-10 Polity democracy scale, else 0	Polity5 Project
	Major power	1 if one state is classified a major power, else 0	CoW State System Membership dataset, v2016
Security			
	Alliance (broad)	1 if a any military alliance exists on the dyad, else 0	CoW Formal Alliances v4.1
	Alliance (defence pact)	1 if a defence pact exists on the dyad, else 0	CoW Formal Alliances v4.1
	Nuclear weapons (both)	1 if both dyad states possess nuclear weapons, and 0 otherwise	Nuclear Weapons – Our World in Data
	Nuclear weapons (one)	1 if only one dyad state possesses nuclear weapons, and 0 otherwise	Nuclear Weapons – Our World in Data
	Dyadic peace years	The number of years since the dyad was last involved in an MID	CoW MID Dataset
Geographical			
	Contiguity	1 if contiguous on land, or separated by less than 24 miles of water, 0 otherwise	CoW Direct Contiguity v3.2
	Distance	Distance between the dyad capitals in kilometres	EUGene
	Regions	1 when both states are in the same region, and 0 otherwise	World Bank

Measured with Pearson's correlation coefficient (Pearson's r), the structural variables sampled in the modelling dataset show quite some correlation between them, as is shown in the matrix below. This will create difficulties for using linear models for prediction, as is explained further in the Modelling and prediction section. Nevertheless, these correlations are expected due to the concepts the variables reflect.

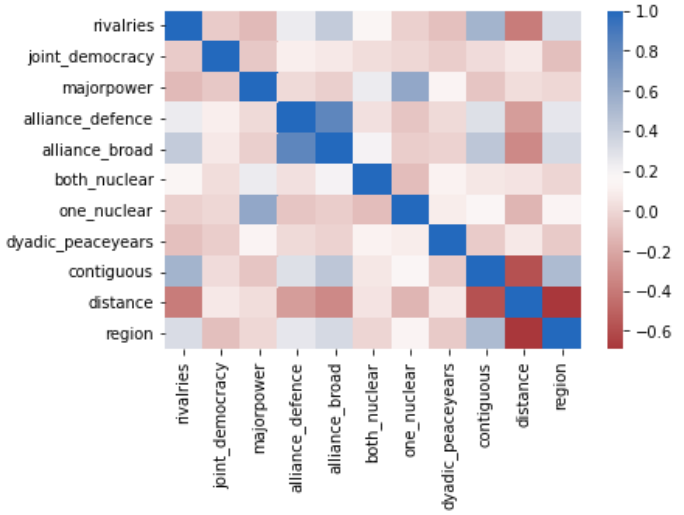


Figure 3-3. Correlation matrix structural variables (Pearson's r)

First, the alliance variables are positively correlated. This is due to the fact that the broad alliance variable is essentially a container for the narrow alliance variable. All alliances that are defence alliances also belong to the category any type of alliance. Second, the contiguity and distance between countries are negatively correlated, as well as distance and being in the same region. This is because all three in essence measure the concept of geographical proximity, but by focussing on different aspects. It is therefore not surprising that they are related. Furthermore, the cluster of geographic variables shows some weak correlation with the alliance and rivalry variables. This also is not surprising. The farther two countries are apart, the less incentive they have to cooperate against a common foe, and also the less interactions there are that could make for a long-standing rivalry. Fourth, possessing nuclear weapons is moderately correlated with being a major power, which is logical, since it is great powers that can afford, and could afford in the past, the costs of development and maintenance of nuclear weapons, and that possess the technical capabilities for such an endeavour. Conversely, possessing nuclear weapons increases a state's power, contributing to a major power status. It is also due to non-proliferation of nuclear weapons, which has prevented non-major powers from acquiring nuclear weapons.

A note on dyadic operationalisation

A key component of the research design is the dyad month as the unit of analysis. For monadic variables, such as economic size, or the number of military conflicts in a nation's recent history, this means that they must be formulated alternatively.

Sometimes, a monadic variable can be converted to a dyadic format with ease, e.g., the presence of a major power in a country-pair. Sometimes however, it is difficult to operationalise a monadic variable dyadically, without losing information regarding the theorised relation to conflict onset and escalation, or even to operationalise it at all. For example, when we want to account for the possible effect a nation's wealth has on its conflict behaviour, we could use GDP as a continuous variable in a monadic setting, but would have to resort to an artificial construct in a dyadic format, such as taking only the highest or lowest GDP of the pair, or averaging them out, losing information in the process.

3.1.2.3 Automatic event data

The other stream of input data is automated event data. Event data are data that captures something happening somewhere in the world, such as a protest, or an election. These data can be manually collected, as has been the case for many conflict datasets, for example the COPDAB dataset (Azar, 1980). However, due to internet news articles, and advances in natural language processing, some projects have automated the collection process of these data, resulting in large swaths of data capturing things happening around the world. These data are called automatic event data. The main goal of these data programs is to capture social dynamics.

At this point in time, the most prominent AED projects are ICEWS and the Global Database of Events, Language, and Tone (GDELT). ICEWS was created as an early-warning system for US Pacific Command, but the data now stretch globally global and are available on a weekly basis (O'Brien, 2010). ICEWS data stretch back to 1995.

GDELT is a competitor to ICEWS, and its data stretch back to 1979, giving it an edge over ICEWS. However, it has been revealed that GDELT data were at least in part fraudulently acquired (*Leetaru v. Board Of Trustees of the University Of Illinois, And Guenther*, 2013), which damaged the reputation of the quality of the GDELT project. For this reason, ICEWS is the source used in this project. ICEWS data are available from 1995 onwards, limiting this project's time period to 1995 – 2014.

Operationalisation and data remarks

The ICEWS data, using the CAMEO coding scheme, distinguish 268 types of events, spread across all categories of international interaction. Examples are 'Express intent to cooperate', or 'Criticize or denounce', or even explicitly violent actions such as 'Use conventional military force'. The typical format of these data is in the shape 'actor A does action X to actor B at time T'. These events were filtered for the time period 1995-2014, for dyadic events, excluding purely domestic affairs, and for relevant countries. Subsequently, they were transformed from the original format into a count of events, of a specific category, on a specific dyad month. Aggregation of some sorts of the individual events is a necessity, since the automatically collected and coded data contain many inaccurately coded data points (Arva et al., 2013; Douglass et al., 2022).

After filtering, 6,233,223 of the original 14,272,027 events remain. These are good for events on 12,226 dyads, or 70.3% of all possible dyads. However, when we turn to dyad months, the canvas shows a lot more blank space.

Multiplying all 17391 dyads with 240 months amounts to a total of around 4 million possible dyad months. Of these, only 9.7% contain an event. This is not surprising however, since the vast majority of dyads do not interact meaningfully at all.

These event counts were then lagged by one month, since the purpose of the model is to contribute to early action, and must thus be trained on the relations between the events happening in one month, and the status of conflict a month later. Eventually, in the final modelling dataset, 4749 out of 9736, or 48.8%, of dyad months contain any events. Of these, 4297 also experienced conflict. Here it can be seen that the event data do in fact reflect at least some real-world activity.

Memory issues

The size of some objects in the data manipulation notebooks is substantial and can cause out-of-memory issues. Although careful choice and design of computations could partially mitigate speed and memory issues, the largest object comes in at around 8.5 GB of memory space, requiring a prohibitive amount of memory space for various computations. This object is a dataframe containing all dyad months (around 4 million), and all assigned event counts (269). To overcome this, Dask was used. Dask is a Python library built on Pandas and offers a similar API. Unlike Python, it allows dataframes to be partitioned and live on-disk, instead of in-memory. Sometimes, this results in longer operation times. However, Dask also offers parallel processing, so when available, this enables significant speed-ups compared to Pandas. The main reason for using Dask however, is its ability to handle data larger than memory.

Using Dask, combined with a 12-node server with approximately 79 GB of free memory space, was enough for the necessary computations in data manipulation and modelling.

3.1.3 Modelling dataset

The final modelling dataset is undersampled from the entire population of dyad months, so that the modelling dataset has a 1:1 ratio of cases of conflict and peace. This is to prevent severe class imbalance, and to improve performance. Nevertheless, substantial class imbalances remain for the onset and escalation target variables. The result is a modelling dataset of 9736 rows and 279 features, divided over 11 structural features, and 268 event count features.

Table 3-4. Distribution of classes, modelling dataset

Occurrence		*Onset*		*Escalation*	
0	4868	0	8906	0 (no MID)	4848
1	4868	1	830	1 (threat to use force)	153
				2 (display of force)	1258
				3 (one-sided use of force)	1280
				4 (reciprocated use of force)	1813
				5 (interstate war)	364
Total	9736		9736		9736

3.1.3.1 Related features

Whether or not there are relations among the features influences the choice for algorithms, mostly because it complicates interpretability. When variables are related, interpreting the specific contribution of a feature becomes more difficult, since features can share an effect on the target variable. How to deal with related features in interpretation can depend on the algorithm used, therefore we examine whether there are relations in the feature space, and whether they should influence the model selection.

There are many different approaches to measure relations between variables. One family of approaches consists of pairwise correlations. Within this group, there are different measures, and which one is most suitable depends on the variables' data type. Since the overwhelming majority of the features is continuous (270 out of 279), and checking assumptions for linearity and homoscedasticity for all pairs is unfeasible, Spearman's rank correlation (ρ) is used, which does not depend on those assumptions. Spearman's ρ measures the strength and direction of a monotonic

relation. In a monotonic relation, when one variable changes, the other variable changes in the same or opposite direction, but the direction must remain constant. To sum up, Spearman's pairwise rank correlation is used to indicate the presence or absence of inter-feature relations.

Figure 3-4 presents the correlation matrix for all features. The measure varies between -1 and 1. In the positive case, it is considered weak between 0 and 0.39, moderate between 0.4 and 0.59, strong between 0.6 and 0.79, and very strong between 0.8 and 1, and in the negative case, vice versa. Due to the size of the feature space, only around a third of the labels are printed, and for the event features, not their names, but their CAMEO codes are printed, so to prevent lengthy descriptive labels. Finally, the dull grey stripes refer to the nineteen features that do not contain any event counts in the undersampled dataset, which is the one used for modelling.

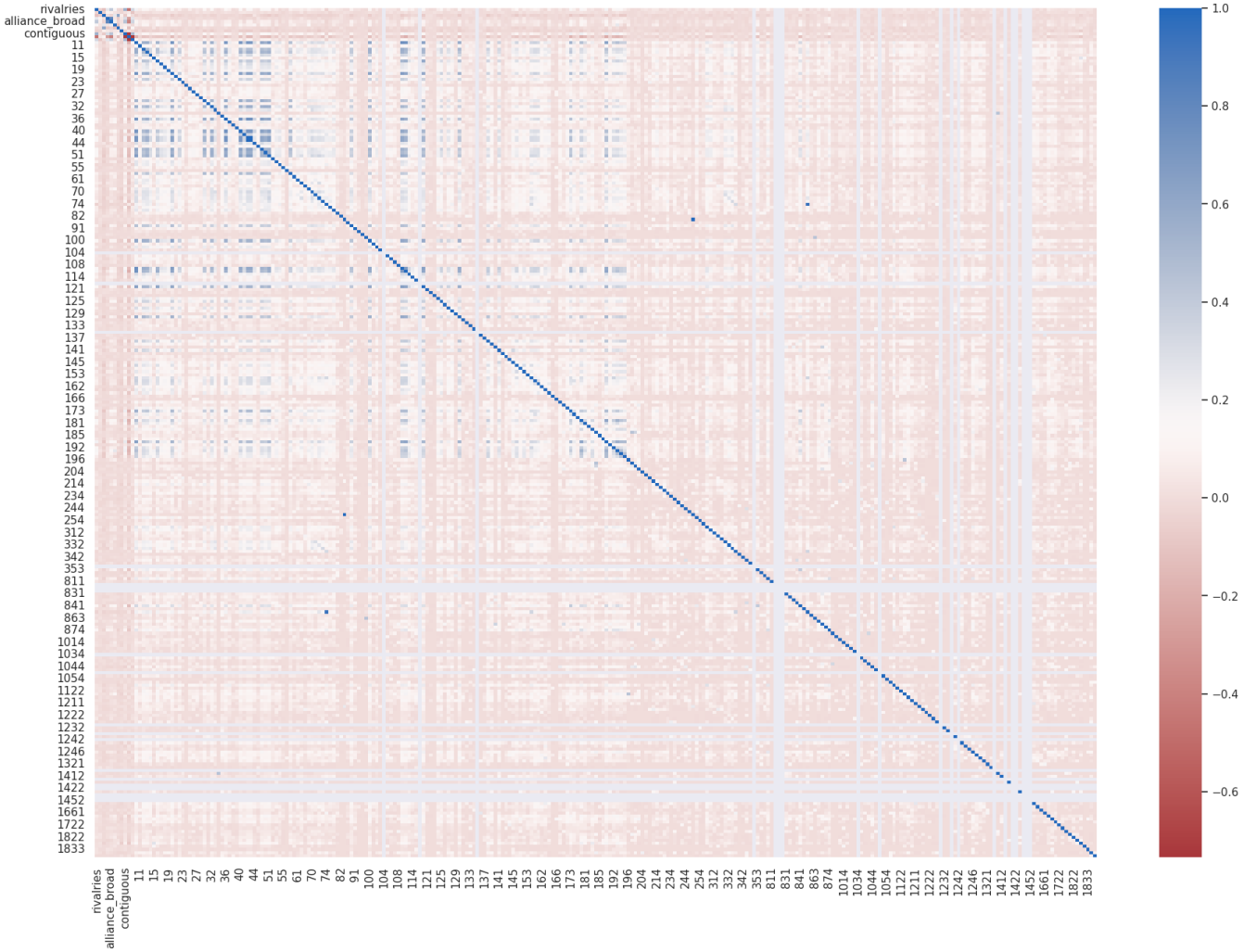


Figure 3-4. Correlation matrix (Spearman's ρ)

In the matrix, a few observations are noteworthy. First, one can clearly recognise the eleven structural variables showing some correlation among each other in the upper-left corner, as in Figure 3-3. Secondly, there are several smaller clusters, or spots, of correlated event count features. This is not surprising, since the events are thematically clustered in the CAMEO coding scheme (Schrodt, 2012). For example, the codes 40 to 46 all refer to some type of consulting between two parties, such as 'Discuss by telephone' (41), or 'Meet at a third location' (44). It seems likely that some event types refer to same processes taking place, and are thus correlated. Finally, we can see some individual bright spots, indicating strong correlations, for example around 860 and around 80. In conclusion, while the overwhelming majority of features does not show strong correlations, there are some bivariate correlations.

The correlation matrix above presents all bivariate correlations. However, since we can identify several smaller clusters in the matrix, it is likely that there are correlations between multiple features as well. This condition is called multicollinearity. Because of the clusters, and since it is good practice, we test for multicollinearity. We do this by calculating the variance inflation score (VIF) for each feature. The VIF is calculated by regressing a feature against the other features. The VIF is then $\frac{1}{1-R^2}$, where R^2 is the R-squared score of the ordinary least squares regression using all other features as input. The VIF scores vary between 1, which indicates no correlation with the rest of the features, to infinity, which signals perfect correlation. Any value above 5 is an indication that

multicollinearity is present, and above 10 is definitely problematic. Finally, unlike Spearman's ρ, the VIF scores rely on linear regression, so it might miss related features that have nonlinear relations. This is not a grave problem, however, since the goal is not statistical scrutiny of the input variables, only to assess whether inter-feature relations exist.

Table 3-5 present the ten highest VIF scores. Indeed, we can see that a decent number of features show multicollinearity, with the top eight showing near-perfect collinearity with the rest of the feature space. In total, 37 features have a higher multicollinearity than 5, and 15 have a higher score than 10. The entire list of VIF scores higher than 5 is presented in Appendix A. Data.

Table 3-5. Top ten VIF scores

CAMEO number	*Feature*	*Variance inflation factor*
252	Appeal for easing of political dissent	∞
83	Accede to requests or demands for political reform	∞
43	Host a visit	20620.82
42	Make a visit	20586.02
74	Provide military protection or peacekeeping	390.01
861	Receive deployment of peacekeepers	389.97
872	Ease military blockade	341.84
140	Engage in political dissent	316.07
190	Use conventional military force	32.27
10	Make statement	19.51

In conclusion, we can state with confidence that there are relations between the features. Both the bivariate correlations, as well as the VIF scores show this

3.1.4 Summary data subsection

The data subsection presented the argumentation behind the most influential choices in data collection and data preparation. The dyad month is the unit of analysis, and the features are operationalised dyadically. Furthermore, both structural variables and event counts are used to predict on three different target variables, conflict occurrence, onset and escalation. The former two are binary, and the latter is categorical. The final classes are balanced for occurrence, severely imbalanced for onset, and quite imbalanced for conflict escalation. Additionally, there are correlations and multicollinearity in the feature space.

For further details, the reader is referred to Appendix A. Data. This appendix presents the logic behind the choice for the CoW MID dataset as the unit of analysis, as well as an in-depth overview of the theoretical links between the target variables and the input variables. The appendix also explains the operationalisation of the input variables in further detail, and provides a list of VIF scores higher than 5.

3.2 Modelling and prediction

The modelling task is a binary classification problem for occurrence and onset, and a multiclass classification for the ordinal escalation target variable. Classification means that models trained on the data predict the classes of a new data point by evaluating whether the calculated probability is above or under a threshold for class inclusion. The next subsection will explain the demands classification algorithms must meet to be of use. Subsequently, the machine learning pipeline is detailed in the following subsection. Finally, the criteria on which the models are evaluated are presented in the last subsection.

3.2.1 Model selection

The algorithms to be used must meet some project-specific demands.

i. The model results must be interpretable. It must be clear to some extent which features are important.
ii. The algorithm must be able to predict well on nonlinear relations. It is unknown whether these exist between the features and the target variables, but due to the size of the feature space (278), it is to be expected.
iii. The algorithm must be able to handle a high dimensionality. The feature space (278) is rather large for the number of cases (9736).
iv. The algorithm must be able to handle imbalanced data.

Demands iii and iv lead us to algorithms that are good at finding nonlinearities in large feature spaces. This is a typical strength of tree-based ensemble methods, but not of linear models. Furthermore, correlation between features does not invalidate the interpretability of tree-based ensemble model results as much as in linear models[10]. Demand iv rejects the use of KNN-like methods, and demand i for interpretability rejects the use of artificial neural nets, which would make a good fit were it not for this demand. In conclusion, tree-based ensemble methods are the best for this task. Within this category, random forest and eXtreme Gradient Boosting (XGBoost) provide two state-of-the-art methods that are well-implemented in machine learning for the Python language. These are the algorithms that will be used.

3.2.1.1 Random forest

Random forests (RF) are an ensemble technique based on decision trees. It was formalised in an influential paper by Breiman (Breiman, 2001) and has since become a very popular algorithm, used across many different fields (Kamiri & Mariga, 2021).

In a random forest, individual trees are built on different subsets of the training data. These subsets are created by bootstrapping the data, which is taking random samples with replacement. Additionally, each tree is built on only a random subset of features. This is to prevent a tree from using the same variable to make splits over and over again. Subsequently, the separate trees each make a prediction on a new data point. By way of majority voting, the forest of trees decides on the prediction. The combination of bootstrapping and aggregating votes is known as bagging.

[10] The fact that linear models have such limitations does not mean they have not been extensively used in the field of quantitative political studies (Schrodt, 2014).

3.2.1.2 XGBoost

Similar to random forest, XGBoost is a tree ensemble method, although the algorithms do have differences, as will be explained further down this section. XGBoost stands for Extreme Gradient Boosting, and is an implementation of the gradient boosting algorithm. The implementation is designed to be efficient and flexible, which greatly contributes to its popularity. It is more recent than random forest, and has become the learner of choice in a variety of applications (Chen & Guestrin, 2016).

The gradient boosting algorithm, also known as gradient tree boosting, is a form of boosting. In boosting, a combination of sequential weak learners is used to make a prediction. A weak learner is a model that is only slightly more accurate than random chance (Rokach, 2009, p. 21). The first successful boosting algorithm was AdaBoost, which was introduced by Freund et al. (Freund & Schapire, 1996) and Freund and Schapire (1997). The algorithm uses decision trees as weak learners. Each tree adapts to the mistakes made by the tree before it. It does so by assigning more weight to the samples that were misclassified, either by changing the training sample weights, or by resampling the training data. The final prediction is then made by combining the separate predictions of all trees, where the more accurate trees get more say in the final prediction than the trees that are performing worse.

Gradient boosting is a generalisation of the AdaBoost algorithm invented by Friedman et al. (2000) and Friedman (2001). Whereas AdaBoost fits its next learner on the tweaked training dataset, gradient boosting fits its next learner on the gradient of the error of the previous predictions, a procedure called gradient descent. Gradient boosting refines its predictions one step at a time, until it approximates the minimum of the loss function. The result is that, whereas AdaBoost is configured with a specific exponential loss function, gradient boosting can use any differentiable loss function, making the algorithm more robust.

The XGBoost library is an implementation of the gradient boosting algorithm described above. It is optimised for efficiency and scalability. One example is using weighted quantiles to determine candidate split points for the tree nodes. Another example would be the possibility for out-of-core and parallel computing. For a more detailed description of the workings of the gradient boosting implementation in the XGBoost system, see the paper by Chen and Guestrin (2016), and the documentation at xgboost.readthedocs.io (2022).

3.2.2 Modelling pipeline

The models were applied to the dataset in four steps. First, the dataset was split in a training and a test set, then the model's optimal hyperparameters were tuned by cross-validation. Third, the model is fit to the training data, and, finally, the model's predictions are evaluated on the test set. All these steps were programmed in the Python language using the SKLearn and XGBoost libraries.

This workflow was repeated twelve times. All three problems were run with both algorithms, once with the structural features only, and once with all features combined. The details of the pipeline can be found in the section below.

Subsection 3.2.2 presents the choices made in the modelling workflow up until the evaluation metrics, which are presented in subsection 3.2.3.

3.2.2.1 Train/test split

The goal of predictive modelling is always to make accurate predictions on new data, in which we do not know the outcome of the target variable. However, this constitutes a difficult problem, since these data are never in possession of the modeller. A standard way to recreate new, real-world cases is to hold out part of the dataset for testing, called the test set. The model is then trained on the training set, and evaluated on the test set.

The data were thus split in a training and a test set in a 0.67 to 0.33 ratio. Since the specific split of the data influences the model training, we tested for sensitivity to the training/test split, which is shown in Chapter 4.

3.2.2.2 Hyperparameter tuning

Finding the optimal configuration

Almost all machine learning algorithms have various parameters that can change the learning process of the algorithm. One can, for example, adjust the number of trees, and the depth of trees in decision trees. These

parameters are called hyperparameters. To optimise performance, one must find the optimal combination of hyperparameters, by testing various configurations on unseen data. However, this may introduce a problem called information leakage. By testing hyperparameter combinations on the held-out test set and selecting the best-performing combination, we would rely on information from the test set to eventually train our model. This information leak reduces the test set's capability to be an objective test of model performance. To counter this problem, one might hold out another part of the training set. On this validation set, one could then test for the optimal hyperparameter combination. This would, however, further reduce the size of the valuable training set.

The solution is a procedure called k-fold cross-validation. In cross-validation, the data are split into folds, with k being the number of folds. Each fold will be used once as a held-out test set, and k-1 times as part of the training set. The model will thus be trained k times, and its average score across the k training times will be the resulting score. This way, we achieve a performance indication of a specific configuration of hyperparameter values, and do not waste any training data in the process.

Subsequently, another advantage of cross-validation vis-à-vis a validation set, is that by using all data in the training set, the selection of hyperparameters is not dependent on the train/validation split. Finally, the value of k is typically 5 or 10. A higher value may result in better estimates of model performance, but also in significant higher computational loads. 10-fold cross-validation did not increase model performance above 5-fold, so k = 5 was used. The procedure for cross-validation is visualised in Figure 3-5.

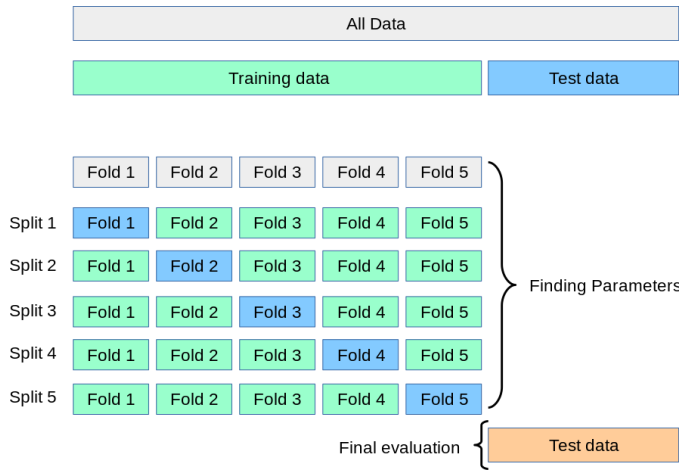


Figure 3-5. 5-fold cross-validation (scikit-learn, n.d.)

Using k-fold cross-validation, we can test a specific hyperparameter configuration. But we still need to test various configurations to find the optimal one. One can do this manually by repeating the cross-validation procedure, but also automatically. The best method to do so is by searching through the entire hyperparameter space (exhaustive search) for the best configuration. This is computationally very expensive, however. Instead, we use a common alternative, randomised grid search, where random configurations are sampled from a pre-defined grid of hyperparameter values, which are then tested in cross-validation. The entire method can thus be summarised as randomised grid search k-fold cross-validation. The tuning grid is detailed at the end of this subsection, because first, we should delve into the issue of under and overfitting to the data.

Dealing with overfitting

Every machine learning modelling effort has to consider the effects of the bias-variance trade-off. A highly biased model contains systematic errors and will not match the training dataset closely. As a result, it is not very sensitive to the data inputs and its predictions may be oversimplified. On the other end of the continuum are high-variance models. These models match the training set too closely, and as such will not generalise well to unseen data. A good model strives to reduce both bias and variance, but will ultimately hit the trade-off. In the former case described, one would say that a model underfits to the training data, while in the latter case, a model is said to overfit.

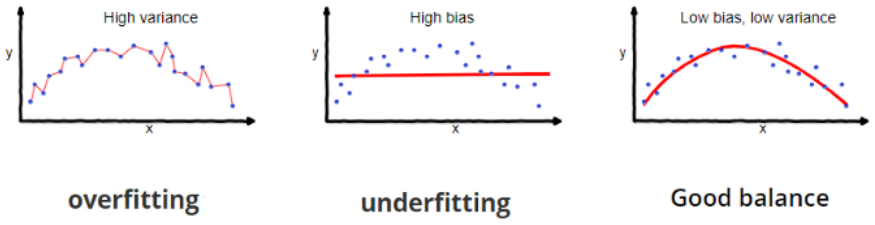


Figure 3-6 Bias-variance trade-off (Singh, 2018)

Decision trees, the basis of both random forest and XGBoost, are known to be prone to overfitting (Schaffer, 1993). Although the bagging and boosting procedures employed in the algorithms do reduce this risk, additional measures were taken by including regularisation parameters in the tuning grids. By comparing different configurations with cross-validation, the optimal values for the regularisation parameters are selected, thus reducing overfitting automatically.

Tuning grids

The tuning grids of random forest and XGBoost contain their hyperparameters and their possible values. This section details what the hyperparameters represent, why they prevent overfitting in case of regularisation parameters, and how their grid values have been chosen.

For random forest, 1) the number of estimators refers to how many trees are used in the ensemble. The more trees, the more accurate the model is, but the more expensive it is to train. 2) The maximum tree depth limits the number of levels a tree can grow. The more levels, the more it fits to the training data. 3) The maximum number of features hyperparameter chooses the metric which indicates how many features may be sampled in each bootstrapped data sample. Since random forest uses feature subsampling, a decision rule must be established to decide how many features should be subsampled at maximum in a bootstrapped sample. This decision rule is either the square root of the total number of features, or the log base 2 of this number. Then, 4) the split criterion used for measuring information gain can be Gini-impurity, entropy, or log-loss. Next, 5) the minimum number of samples per node places a minimum on a node to be further split. The higher the minimal node sample number, the less the trees are allowed to overfit. 6) The minimum impurity decrease required refers to the information gain splitting a node must provide. The higher the minimal decrease, the more restricted the algorithm is in overfitting to the training data. Finally, 7) scikit-learn offers the possibility to turn off bootstrapping, keeping only random feature selection as a source of variance for the data samples for the trees to grow on, if this may help the model's performance.

Table 3-6. Hyperparameter tuning grid - random forest

Hyperparameter	*Possible values*
No. of estimators	200, 400, 600, 800, 1000, 1500, 2000
Maximum tree depth	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, None
Maximum no. of features	Square root, $\log_2$
Split criterion used	Gini, entropy, log_loss
Minimum no. of samples to further split a node	2, 3, 4, 5, 6, 7, 8, 9, 10
Minimum impurity decrease required for split	0.0, 0.05, 0.1
Bootstrapping	True, False

The first two hyperparameters in XGBoost, number of estimators and maximum tree depth, are similar to random forest. The last two hyperparameters can also be compared to a counterpart in random forest, but are implemented somewhat differently. First, gamma is comparable to the minimum impurity decrease. It specifies the minimum reduction in the loss function required for a split. The higher gamma, the more restrained XGBoost is. Second, the minimal child weight is comparable to the minimum number of samples in a node in random forest. However, it measures the sum of sample weights in a child node, rather than the number of samples in a potential parent node. The sum of sample weights is a better choice in the context of XGBoost, since gradient boosting assigns weights to individual samples, and uses those weights for the next tree. The higher the minimal child weight is, the less the algorithm fits to the training data.

Consequently, the other three hyperparameters are only applicable to XGBoost. First, the learning rate, or shrinkage factor, reduces the effect of new trees added to the model. The corrections proposed by a new tree are scaled down by the learning rate, effectively slowing down the learning process. The smaller the learning rate, the less likely the model is to overfit, although it is much more computationally expensive to train.

Then, the last hyperparameters are the alpha and lambda, who respectively represent L1 and L2 regularisation. Both are methods to reduce the weights of the leaves of the weak learners. They introduce an extra error term into the loss function, penalising complexity and preventing the model from overfitting. Alpha is strict in its approach, often reducing the weights to zero, and thus enforcing sparsity. L2 regularization is a softer approach, reducing the weights, but without preference to reduce them to zero. Higher values translate to a more conservative model for both alpha and lambda, meaning that the model is less prone to overfitting.

Table 3-7. Hyperparameter tuning grid – XGBoost

Hyperparameter	*Possible values*
No. of estimators	50, 100, 150, 200, 1000
Maximum tree depth	3, 4, 5
Learning rate	0.01, 0.05, 0.1, 0.2, 0.35, 0.5, 0.7
Alpha	0, 0.1, 0.4, 0.8 ,1.6, 6.4, 25.6, 102.4, 200
Lambda	0, 0.1, 0.4, 0.8 ,1.6, 6.4, 25.6, 102.4, 200
Gamma	0, 0.5, 1, 1.5, 2, 5
Minimal child weight	1, 5, 10

Finally, for both algorithms, some hyperparameters have a limited number of values. Bootstrapping, for example, can be either be used or not, without any other options. Others have infinitely many different values, such as the number of estimators. In this case, we checked whether the grid ranges were wide enough to see whether the grid search was being constrained unnecessarily. When the optimal value for a hyperparameter was located at the end of the range, we checked whether expanding the range increased the cross-validation scores. Except for the maximum tree depth of XGBoost, this was not the case. However, higher XGBoost tree depths increased overfitting, and were thus limited to 5 at the highest.

3.2.3 Model evaluation

Model evaluation is a critical component of any machine learning project. The subsection below details what metrics were used to evaluate the different models, and sets out why these metrics are relevant to the problems at hand. It does so first for the occurrence and onset problems, and then for escalation as well, which as a multiclass problem requires different metrics. Finally, it describes the concept of feature importance, and how it is used in this study.

3.2.3.1 Binary classification metrics

For binary classification, the confusion matrix is the starting point, and additionally is the basis of most evaluation metrics. It provides an overview of the model's predictions versus their actual value. Below, the metrics used in this project are set out.

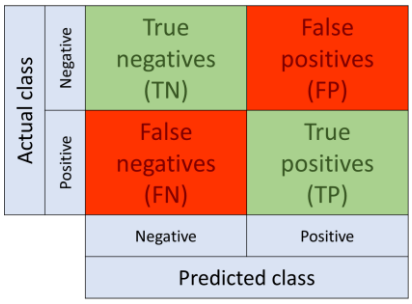


Figure 3-7. Confusion matrix

F_{beta}-score

The F_{beta}-score is the most important evaluation metric for the binary classification task. It is a generalisation of the popular F1-score, which is the harmonic mean between the precision and the recall. Precision is defined as the ratio of true positives to all predicted positives, and recall as the ratio of true positives to all actual positives. Since we are mostly interested in false negatives, recall is the more important aspect of the F_{beta}-score.

Instead of the F1-score, which is already useful since it only focuses on prediction of the positive class, in our case being dyad months experiencing conflict, the F_{beta}-score can be customised to put more weight on the recall relative to the precision. This is done by varying the beta value: a higher score increases the recall's importance. In this project, the chosen beta is 2, signifying that not missing any conflict cases is substantially more important than mistakenly assigning a conflict score to peaceful instances. Finally, similar to the accuracy, the F_{beta}-score uses predicted class scores, not probabilities, and is therefore dependent on the probability threshold chosen for classifying a conflict as either 0 or 1. In this study, the default threshold value of 0.5 is used. While this is the most intuitive (a probability higher than 0.5 points to conflict, lower than 0.5 to peace), it might be possible to increase performance by testing different classification thresholds in cross-validation, like the model hyperparameters.

$$F_{beta} = \left(1+\beta^2\right)\frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

Average precision (AP)

Although the F_{beta}-score evaluates the models performance on the false negatives in relation to the positive class in general, we are also interested in how well the model performs across all probability thresholds. Therefore, we use the precision-recall curve to evaluate how well the model predicts conflict in general. The average precision (AP) is the area-under-the-curve of the precision-recall curve and captures the model's performance in a single metric, precision-recall curve area-under-the-curve (PRC AUC), or average precision (AP). But, since false negatives are the costliest misclassification, the F_{beta}-score is more important when the average precision and F_{beta}-score are not in agreement.

Receiver operating characteristic area-under-the-curve (ROC AUC)

The ROC is a curve that evaluates sensitivity and specificity across probability thresholds, and is one of the most popular metrics to evaluate classification models in the field of conflict prediction, as well as outside of it. However, it is not the most appropriate score in our case, since it cannot make a distinction between how well the model predicts the positive and the negative class, and neither on the difference in cost of false negatives, and false positives. Additionally, the ROC AUC is not suited for imbalanced class problems. The current data setup allows for a balanced class structure, but once we would like to test for imbalanced structure, this score will not be useful. Consequently, the main contribution of this score is to create some comparability with other studies.

Matthews correlation coefficient (MCC)

The MCC can be considered the best score of performance across all four confusion matrix categories, especially in imbalanced datasets, as it accounts for class size. We will use this score to gain an idea of general performance, instead of the ROC AUC, while noting that general performance is of secondary importance.

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3.2.3.2 Multiclass classification metrics

The confusion matrix for multiclass classification is less clear-cut than for the binary case. No conflict still is the negative class, and the various measures of escalation form the positive class. Within the misclassifications for these measures, it is not so clear what constitutes a false negative or a false positive. Instead, they are over, or under-classified. In any case, the goal here is to get the level of escalation correct, and it is less important whether the model over, or under-classified. This new matrix has some implications for the usage of metrics.

Actual class							
	0	TN	FP	FP	FP	FP	FP
	1	FN	TP	x	x	x	x
	2	FN	x	TP	X	x	x
	3	FN	x	x	TP	x	x
	4	FN	x	x	x	TP	x
	5	FN	x	x	x	x	TP
		0	1	2	3	4	5
		Predicted class					

Figure 3-8. Confusion matrix multiclass problem

First, instead of the F_{beta}-score with beta = 2, we use beta = 1, in effect using the F1-score. This is because, in contrast to the binary case of onset and occurrence, we are now interested in all misclassifications, not only the false negatives. In this setting the F1 score is calculated per class. To average the class score into one metric, we use macro averaging, meaning that each class holds the same importance, as opposed to micro or weighted averaging, where the final metric would be heavily influenced by the performance of the zero class, since this class by far has the most cases.

Second, the average precision and the ROC AUC are not very intuitive in a multiclass setting, and are therefore not used. Specifically the ROC AUC is not very helpful, since it cannot be used to compare with other, binary, studies, which was the reason for including it in the previous tasks. Therefore, it is replaced by the MCC, which is also suitable for multiclass classification, and represents the performance of the model across all classes.

3.2.3.3 Feature importance

The last step in model evaluation is not related to the validity of the model itself, but to the importance of the features used. This is to gain insight into which factors are important in predicting conflict, as stated in sub question four. To assess how useful features have been in model training, the feature importance scores are calculated. There are different ways of arriving at feature importance measures. In this study, the mean decrease in impurity (MDI) scores are used. The following subsection describes what these feature importances are, how they should be interpreted, and compares them to similar measures.

In general, feature importance scores evaluate the contribution of individual features to the model's prediction. This is different than the coefficients in a method like linear regression. If all statistical assumptions are met, regression coefficients can be used to make statements about the true data generating mechanism, for example by examining statistical significance. In contrast, the tree ensemble methods used here do not impose any statistical demands on the data. However, their feature importance scores should rather be seen as quantifications of how much a feature played a role in acquiring the prediction accuracy (Debeer & Strobl, 2020). Following these characteristics, feature importances indicate which variables are important for prediction, and which may be correlated or causally important as well, but this cannot be deduced from feature importances alone.

The MDI scores used here were originally introduced by Breiman alongside the introduction of the random forest algorithm (Breiman, 2003). They measure the decrease in impurity by a split of a given variable, across all splits and all trees. Hence, the name mean decrease in impurity. Since Gini-importance is an oft-used measure of impurity in classification decision trees, MDI is sometimes also called Gini-importance. This score was used for both random forest and XGBoost. In short, the MDI scores indicate how valuable a feature has been for model training. It is computationally inexpensive and offers a look into the black box model that an ensemble of trees otherwise is. Nevertheless, the MDI scores have noteworthy shortcomings.

First, the MDI scores are calculated with the training sample. This can result in overly optimistic scores for features which might be less useful in predicting a sample in new, unseen data than others. Second, the method is biased towards high-cardinality features, i.e., numeric features or categorical features with many classes. This is because those features, in contrast to low-cardinality features, offer many opportunities for splitting the data sample, which is what decision trees do. Thus, they will be overrepresented in the model structure, and disproportionally contribute to impurity decrease, distorting the MDI scores, as shown by Strobl et al. (2007), and further discussed by Boulesteix et al. (2012). Then, the scores are vulnerable for correlations between the features. This effect is more pronounced in random forest than in XGBoost, though. Since random forest uses random feature subsampling, correlated features each show up in aggregations of impurity decreases. As such, they share the responsibility of the underlying predictive phenomenon, thus reducing their importance proportional to the size of the correlated group. This effect was called correlation bias by Toloşi and Lengauer (2011). In XGBoost, however, instead of random subsampling, the next weak learner improves the aggregate function of previous learners. Thus, correlated features do not bring much new information, and are generally discarded during model training. Finally, the MDI scores only provide an estimate of the size of an effect, but to not provide directionality. This, however, is common for many other feature importance techniques as well.

When interpreting MDI scores, one should account for the shortcomings and the character of the technique. In practice, this means that the absolute, or scaled, value is not precise, and that one can best rely on the ordinal feature ranking. Secondly, one has to acknowledge the effects of high-cardinality and correlation bias on the features at hand.

Finally, these shortcomings have long been recognised in the scientific literature, as shown above. Due to the popularity of machine learning models, including tree-based ensembles, and random forest in particular, much attention has been paid to making black box models more interpretable[11]. While this study uses the MDI scores, it might be useful to corroborate these with other measures. A first candidate would be permutation feature importances, also introduced by Breiman (2001). These compute a feature's contribution by noising up that feature, and then assessing the model's performance. Benefits of this approach would be that it can be evaluated on the test set, and would be less affected by high-cardinality features, although the correlation bias might be more serious than with MDI scores. A solution to this problem is proposed by Strobl. et al. (2020), although their conditional permutation feature importance method is as of yet only implemented in R, and not in Python. Another alternative would be to pre-process the data, by selecting only one feature from each cluster of correlated variables, before

[11] For an overview of the state-of-the-art, see Molnar et al. (Molnar, 2022). One key insight is that there is no consensus on a rigorous definition of model interpretability.

computing permutation importances. This would come at the cost of potential information loss of the excluded features however, and is thus inferior to conditional permutation feature importance.

3.3 Conclusions

To predict the phenomenon of interstate conflict, three measures are used, with the undirected dyad month as unit of analysis. First, the occurrence of conflict, as this offers the best prospects for successful prediction due to the relatively large number of positive cases. The second measure is the onset of conflict, which is much rarer than occurrence, since only the actual starting month of any MID in a country pair classifies as conflict onset. Third, escalation can offer a prediction for the severity of conflict taking place. These three concepts of interest are to be predicted with a feature set consisting of eleven structural features, and 268 event count features. The feature space has been shown to contain relations between the features.

Subsequently, each target variable presents a different prediction problem. The first two problems, occurrence and onset, are binary classification problems. Escalation, however, has an ordinal scale, which calls for a multiclass classification method. Due to the large size of the feature space, the need for interpretability, and expected multicollinearity and nonlinearities, random forest and XGBoost are the algorithms selected for the prediction task.

Then, the data are split in a training set and a test set in the ratio 0.67 to 0.33. The training set was used for hyperparameter tuning via randomized grid search cross-validation. In the tuning grid, regularisation parameters were included to prevent overfitting of the models. For the binary task, the grid search evaluates the hyperparameters on the F_{beta}-score, with beta = 2. For the multiclass task, the macro-averaged F1 score fulfils this role. The model is then retrained on the entire training set with the optimal hyperparameters, and its performance is assessed on the test set. Once again, the F_{beta} and F1 score are the principal evaluation metrics, together with the average precision.

Finally, impurity-based feature importance scores are calculated to interpret the contributions of individual features, which, although imperfect, can shed some light of on the inner workings of the models.

4 Results and Analysis

This chapter presents and analyses the results of the predictive models. It is structured in four main parts. The reader is guided through the main results by two questions in the first two sections. First, does including event data improve predictions of the models? This effect is clarified in the first section. Second, if a model has predictive power, what are the main variables that drive it? This is set out in the Feature importances section. Both of these questions refer, in turn, to the three connected but distinct problems this study researches. These are the occurrence, the onset, and the level of escalation of interstate conflict on any given dyad in any given month. Finally, the third section discusses the model's sensitivity to the train/test split and how much the models overfit.

4.1 Effects of including event data

Does the inclusion of event data into prediction models of interstate conflict based on structural values improve the predictive power of these models? That is the question at hand in this section. It is evaluated separately for conflict occurrence, onset, and escalation, in their respective subsections.

For each problem, four different setups were run. Both algorithms were run with, and without added event count features. Since the confusion matrix, and the ROC and PR curves, available for each model run, can be informative when plotted, they would clutter the section when all would be reported. Therefore, in each subsection, the algorithm that performs best on the combined feature space is selected without visual inspection, and only this one is evaluated and presented visually. This model is then compared to the best-performing model on the structural-only feature set, using only the confusion matrix for visualisation.

The included event data features are simple counts of the number of events occurred in that dyad month. To evaluate the contribution of these features, the crux of the comparison is whether the all-feature models score better than the structural-only models.

4.1.1 Occurrence

Occurrence is operationalised dichotomously. With 4896 dyad months of conflict occurrence, and exactly 4896 negative cases, this is the easiest task of the three, since it has the most data samples available for learning. Since it is binary, the metrics used for evaluation are the F2-score, the average precision, ROC AUC, and the Matthews correlation coefficient[12]. Of these, the F2-score is the most important, because it places a heavy emphasis on preventing false negatives. Of secondary importance is the average precision, which places an emphasis on correctly predicting conflict, but without explicit focus on false negatives. Finally, the ROC AUC and MCC evaluate the model's predictions across all categories, which is of less importance than the more focused F2-score and AP.

Below, we will first evaluate performance of the models with all features, before comparing the best model to the best model using only structural features. All evaluations are on the test set, meaning out-of-sample.

When using all features, both models perform similarly, and both perform quite well. The F_{beta}-score is around 0.97, and the average precision hovers around 0.99. These are relatively high values which means that the models can comfortably predict conflict occurrence, without many false negatives. This can also be seen in the confusion matrix, where it is shown that out of 1616 conflict occurrences, random forest misses only 46 cases. Furthermore, the number of false negatives is lower than the number of false positives. This is an indication that tuning the hyperparameters to optimise the F2-score did work to bias the model towards reducing false negatives above false positives. The reason for this approach is that not predicting an actual conflict is more costly than incorrectly predicting one. Conflict can cause huge damage to material resources and human lives, which might have been mitigated if predicted. Incorrectly predicting conflict, on the other hand, will most likely not lead to the same losses, but rather to a waste of resources spent on conflict prevention or mitigation.

The results of both models are presented in the table below. Neither model substantially outperforms the other, although the random forest model seems to have a slight edge. The small differences in scoring, however, might as well be due to the models using this particular training sample as a part of the total data sample. To see the size of this effect in more detail, the reader is referred to the test of sensitivity of the models to the train/test split, presented in subsection 4.3.1.

Table 4-1. Comparison algorithms - Occurrence

Metric	*Random Forest*	*XGBoost*
F2	0.970	0.969
Average precision	0.988	0.988
ROC AUC	0.986	0.989
MCC	0.934	0.930

[12] For their operationalisation, see section 3.2.3.

When inspecting the plots for the PRC AUC (equal to average precision) and the ROC AUC, we can see that the models behave as expected. However, before noting these models' performance, let us be aware that the more the curves are pushed into a corner, the upper-right corner for the PRC AUC, and the upper-left for the ROC AUC, the better the model performs across all probability thresholds and the higher the score is. In agreement with their high scores, this pattern is shown in both the PRC AUC and the ROC AUC plots. Furthermore, the lines for both graphs are smooth, indicating that there are multiple features contributing incrementally, and that there is no anomaly in performance across different thresholds. Finally, we can see that the best model, which is based on the F2-score prefers to achieve high recall above high precision. This confirms that the F2-score works as intended and that it focusses on false negatives.

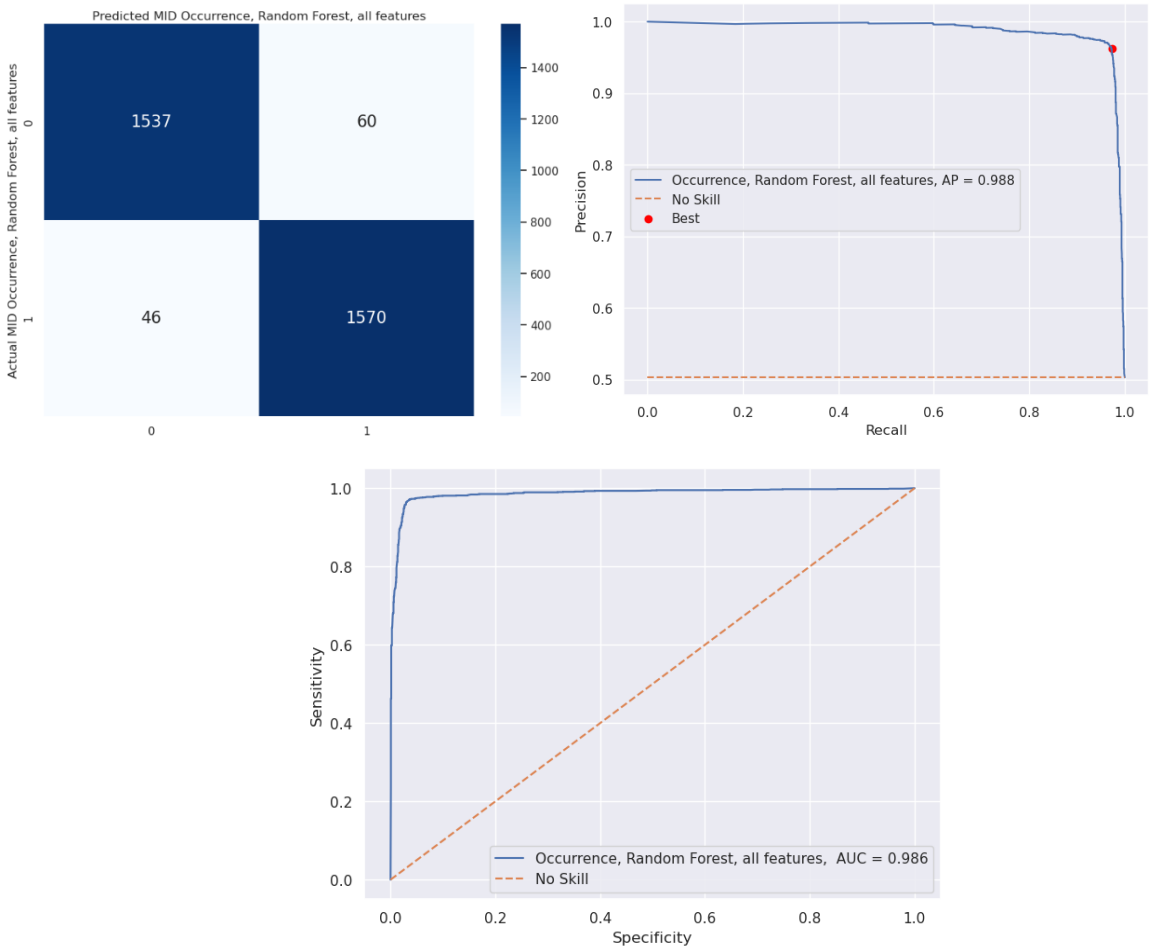


Figure 4-1. Best model results – occurrence

Having established that the models that use all features perform well in predicting conflict occurrence, let us compare them to the scores achieved by the structural models. Table 4-2 presents the performance of the best-performing structural and the best all-features model.

Table 4-2. Comparison all features/structural – occurrence

Metric	*Random forest (structural)*	*Random forest (all features)*
F2	0.970	0.970
Average precision	0.993	0.988
ROC AUC	0.992	0.986
MCC	0.928	0.934

The best algorithm for only structural data was random forest. However, what is interesting, is that the addition of event data does not improve modelling performance. The average precision and ROC AUC even reported a mild decrease. Thus, it is impossible to state that adding event counts improves the prediction of conflict occurrence. The fact that event data counts do not improve performance may be due to the fact that the benchmark of the structural models is already quite high, but is more likely to be due to other factors. In the onset case, presented in the next subsection, the event counts fail to deliver any improvements as well, although the structural model benchmark is much lower for this case.

Further interpreting the structural models, the confusion matrix of the structural models presented in Figure 4-2 shows a pattern that corresponds with the metric outcomes, and which is similar to that of the all-features models. Most cases have been predicted correctly, and the number of false negatives is higher than the number of false positives, showing that the algorithm succeeded in prioritising the prevention of false negatives, or 'missed conflicts', above peace instances classified as conflict.

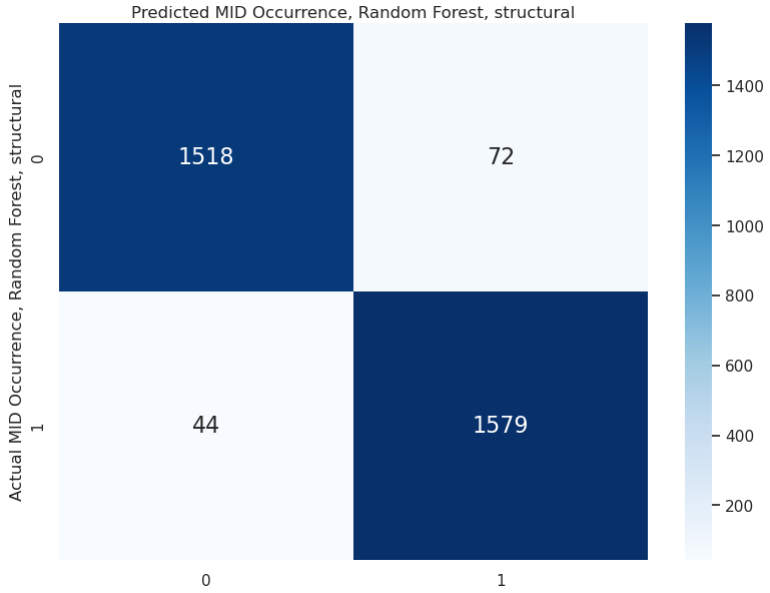


Figure 4-2. Confusion matrix only structural, RF - occurence

The conclusion here is that the structural variables alone are already quite powerful in predicting conflict occurrence. Adding event counts as a large basket of features does not improve the models.

4.1.2 Onset

The second problem to be predicted is conflict onset. The target variable is severely imbalanced, even after undersampling the majority class of non-onset cases. For a total of 9736 cases, there are only 830 onset cases against 8906 non-onset cases. These 830 cases have been split over the training (0.67) and test sets (0.33). The paucity of conflict onsets makes this problem harder to predict than the previous one. As for the evaluation metrics, the logic is largely the same as was detailed above for the evaluation of occurrence. One exception is that, due to the class imbalance, the class-specific metrics (F2, AP) become even more important relative to the general metrics (ROC AUC, MCC) than they already were. Ultimately, any user interested in conflict onset desires to know how well the model predicts conflict onset cases. They would not be interested to know that the model can predict the majority of cases correctly, when that does not provide any information about conflict onsets specifically.

The following paragraphs present and analyse the results of the conflict onset models. It can be stated in advance, however, that none of the setups yield satisfactory results. Both algorithms are unable to identify conflict onset effectively, both with structural features and with the combined feature set.

Table 4-3. Onset prediction results, all features

Metric	*XGBoost (All features)*	*Random forest (All features)*
F2	0.132	0.114
Average precision	0.255	0.241
ROC AUC	0.784	0.748
MCC	0.201	0.161

The table above lists the performance of both models, using all features. In evaluating performance, the F2 metric is the most important. For both XGBoost and random forest, the F2 scores are very low, respectively 0.132 and 0.114. The same goes for the average precision, with values lower than 0.3. The ROC AUC scores are higher, between 0.74 and 0.79. The benchmark score for the ROC AUC, however, is 0.5, which would equate to a coin toss, so these scores indicate that there is some predictive power in the models. However, this is largely due to the automatic assignment of non-onset cases. These form a large majority. This translates into higher ROC AUC scores, but this is not of interest here, because they do not reflect the class of interest.

The distribution of predictions of the most successful all-feature model (XGBoost) is shown below. The model fails to discern onset cases, and overwhelmingly assigns peace values to conflict episodes. The single positive aspect is that around half of the predicted conflict cases are actual conflict onsets. This at least gives the model some limited use as a horizon-scanning tool which misses many onsets, but is 50 per cent sure in the cases it does assign conflict onset. This could be used to alert policy makers to conduct further analysis, but even then, it remains of limited use.

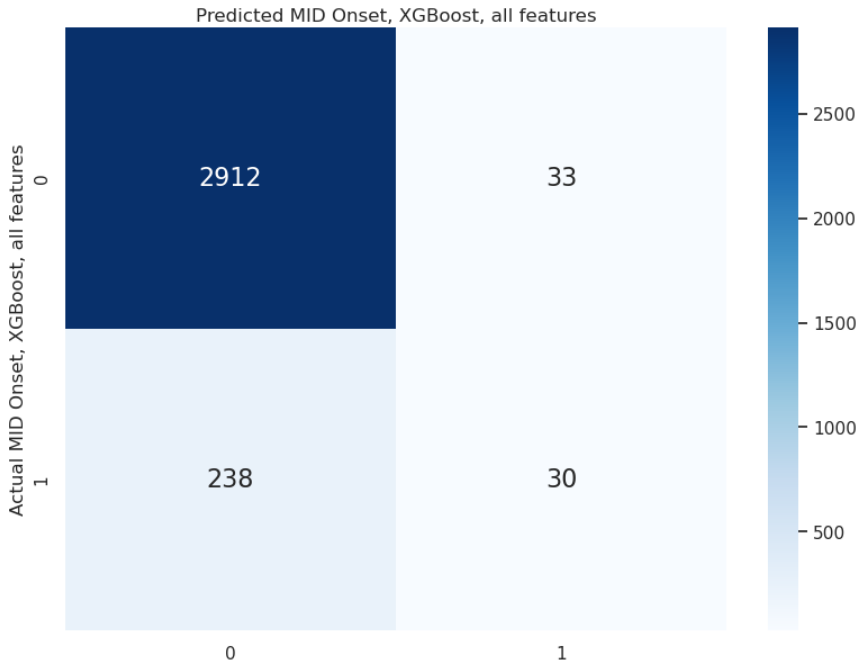


Figure 4-3. Confusion matrix all features, XGBoost - onset

Having concluded that the models using the entire feature space do not have much predictive value, the question now is whether the structural-only models perform worse, indicating a small improvement by the event counts. As we shall see, this is not the case. Table 4-4 below presents the results for the best structural model, alongside the best all-feature model already shown in Table 4-3. The structural model also presents poor results.

Remarkable is, however, that the structural model scores are less dissatisfactory than those of the all-feature model. Event counts seems to slightly worsen performance. This is remarkable, since it runs counter to the expectation that event counts can capture the short-term dynamics that lead up to conflict onset. If anywhere, they could contribute here: the structural model scores are poor, so there is much room for improvement by the event counts. One explanation for the is lacklustre performance of the event count features is that the added features might contain too

many noise features, in effect driving down the predictive power of the effective predictors. Another is that the operationalisation as simple counts might be too straightforward. This will be further discussed in section 5.1.

Table 4-4. Comparison all features/structural

Metric	*Random forest (structural)*	*XGBoost (all features)*
F2	0.254	0.132
Average precision	0.233	0.255
ROC AUC	0.678	0.784
MCC	0.292	0.201

Another explanation for the poor performance across both structural and all-feature models, is the severe class imbalance. However, testing this idea by running the modelling pipelines for conflict onset with ratios of 1:5, 1:3, and even 1:2, instead of the ratios of around 1:9 did not increase performance dramatically, although some gains were made. However, by such severe undersampling, generalisability to the real-world ratio of peace and conflict will come under pressure, since that distribution is very imbalanced (the CoW data used in this study contains a ratio of 1:5010 for monthly onsets in the period 1995-2014).

Finally, some attention must be paid to the degree of overfitting in the all-features models, which is severe. Random forest is most severe, but also XGBoost shows considerably better performance on the training set. Figure 4-4 introduces the in-sample predictions of the XGBoost model. It correctly predicts around forty per cent of the onset cases, and of the cases it predicts as onset, most indeed are an onset case. Meanwhile, the out-of-sample prediction set predicted only around ten per cent of onset cases correctly. This large gap in performance between the training and the test set demonstrates overfitting, meaning that the model does not generalise well. It is good practice to reduce overfitting, for example through further hyperparameter tuning by targeting specific hyperparameters, or perhaps other probability thresholds. However, we should not expect this to drastically increase predictive power. The reason for this is that random forest overfits significantly more than XGBoost, but its predictions are not significantly worse, leading us to expect marginal gains, but not much more, when reducing overfitting.

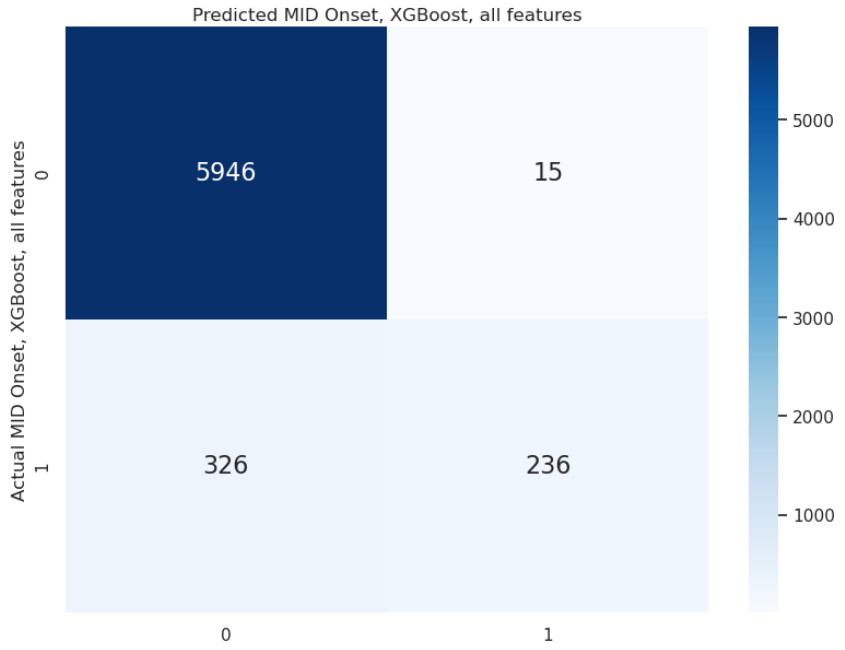


Figure 4-4. In-sample confusion matrix, RF, all features - onset

In conclusion, the performance of both structural and all-feature models is poor for the onset of MIDs on a monthly scale. The models do not succeed in separating conflict onset cases from cases without conflict onset. Including event counts as predictors does not improve performance.

4.1.3 Escalation

The escalation target is an ordinal variable coding for the highest level of escalation experienced during the conflict the dyad month belongs to. It is a multiclass classification problem, which has different metrics than those for binary classification[13]. The most important metric that is changed, is the F2 score, which is now the F1 score. This means that in its formula, recall is equally important to precision, thus not prioritising the reduction of false negatives above the reduction of false positives anymore. This is so, because in predicting escalation, we are not necessarily mainly interested in the instances where the model assigns escalation levels that are too low, but also where the model predicts escalation levels that are too high. Therefore, we use the more balanced F1-score. Subsequently, the other metrics also change somewhat. The ROC AUC and the AP do not make sense in a multiclass setting, since they are meant to be binary metrics. Therefore, we use the MCC to evaluate general predictive power. Finally, precision and recall are also reported for completeness.

The target variable distribution in the total data sample is presented below. Four out of six classes have a decent number of samples, but level 1 and 5 are populated thinly for predictive purposes.

** Excerpt of Table 3-4*

Level of escalation	No. of cases
0 (no MID)	4896
1 (threat to use force)	153
2 (display of force)	1263
3 (one-sided use of force)	1292
4 (reciprocated use of force)	1824
5 (interstate war)	364

Because it is a multiclass problem, both the averaged and separate class scores are reported. The macro-averaged model scores are simply the class scores summed and divided by the number of classes. This does not account for class size, which would skew all scores towards the largest class ('No MID'). The averaged results are presented in

Table 4-5. This time, both the all-features models, and the best structural model are presented in the same table.

Table 4-5. Results prediction escalation, averaged

Metric	*Random forest (all features)*	*XGBoost (all features)*	*Only structural (Random forest)*
F1 (macro-averaged)	0.85	0.87	0.90
MCC	0.86	0.88	0.91

Both algorithms are equally effective when using all features, with a slight edge for the XGBoost model. Similar to onset, the performance actually increases when event data are left out, contrary to the expectation that event data improve performance by capturing dynamic processes. However, with and without event data, the models seem, to an extent, to be capable of predicting the highest level an MID will escalate to.

[13] For details, see subsection 3.2.3.

Below, Table 4-6 presents the prediction scores per class, or level of escalation. The best scores are attained in the 'No MID' class, and the 'Interstate war' class. These results are further discussed on the basis of the confusion matrix in Figure 4-5.

Table 4-6. Results prediction escalation, per class (XGBoost, all features)

Level of escalation	*precision*	*recall*	*F1-score*	*Class size*
0	0.96	0.97	0.97	1597
1	0.64	0.88	0.74	41
2	0.83	0.77	0.80	416
3	0.84	0.81	0.83	445
4	0.90	0.95	0.93	581
5	0.99	0.95	0.97	133
Macro-averaged	0.86	0.89	0.87	3213

When we zoom in on the specifics of the class predictions from the best model using all features, we can see the model does a fairly decent job at separating different levels of escalation, except for category 1 (Threat to use force). Unsurprisingly, this category is by far the smallest of the six. There seems to be a relation between class size and how well the model can predict the class, which is not surprising. What the model does seem to do rather well, however, is separating interstate war from other MIDs, as can be seen at the intersection of predicted and actual data classes for the last class in Figure 4-5. The row and column adjacent to it are almost completely empty.

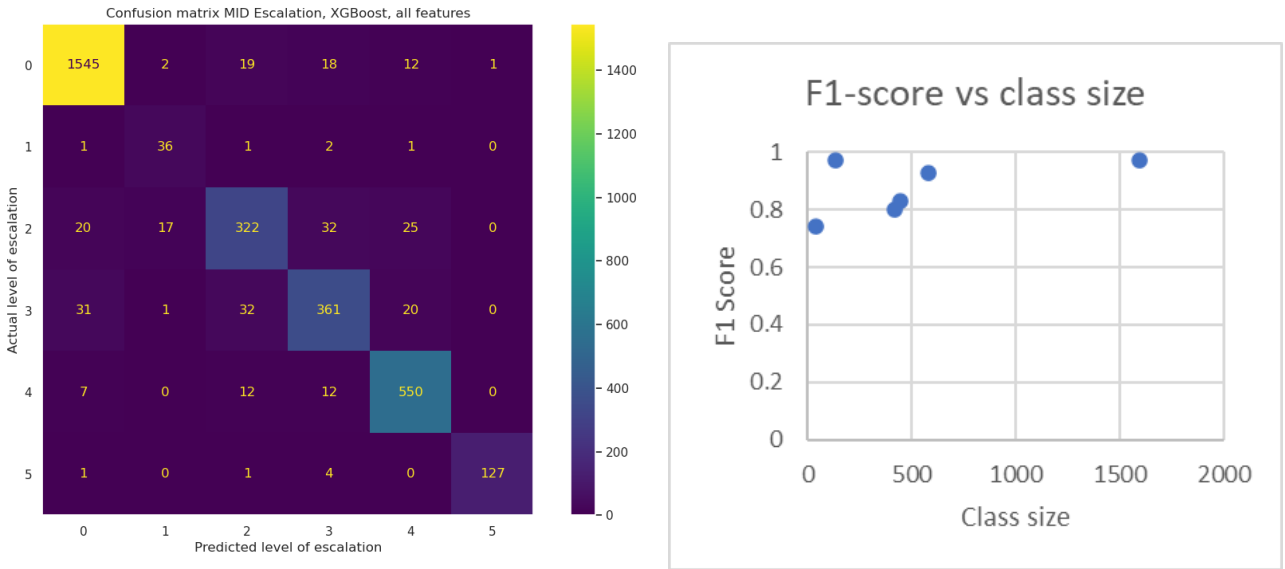


Figure 4-5. Confusion matrix escalation, RF, all features (left), and relation class size/model performance (right)

4.2 Feature importances

Finally, the MDI feature importance scores are given for each algorithm, based on the full feature space. This indicates which variables have been most impactful in the models' predictions, and allows corroborating the results on the inclusion of event data.

4.2.1 Occurrence

Considering the feature importance scores for the prediction of occurrence in the figure and table below, two things are remarkable. Namely, that 1) contiguity seems to be the most important predictor, and that 2) either all (random forest) or most (XGBoost) structural features have higher scores than their event count counterparts.

That the dichotomous contiguity feature in finding 1) was important for model construction is further indicated by the fact that MDI scores are heavily biased towards numeric variables, as is detailed in subsection 3.2.3. Both algorithms using contiguity as a top variable indicates it is a useful feature. The feature itself is a proxy for geographic proximity. Whenever two states share a land border, or are separated by no more than 24 miles of water. The effect of contiguity on conflict occurrence, and on escalation as well, has long been known in the field of conflict studies[14], and is confirmed by this study.

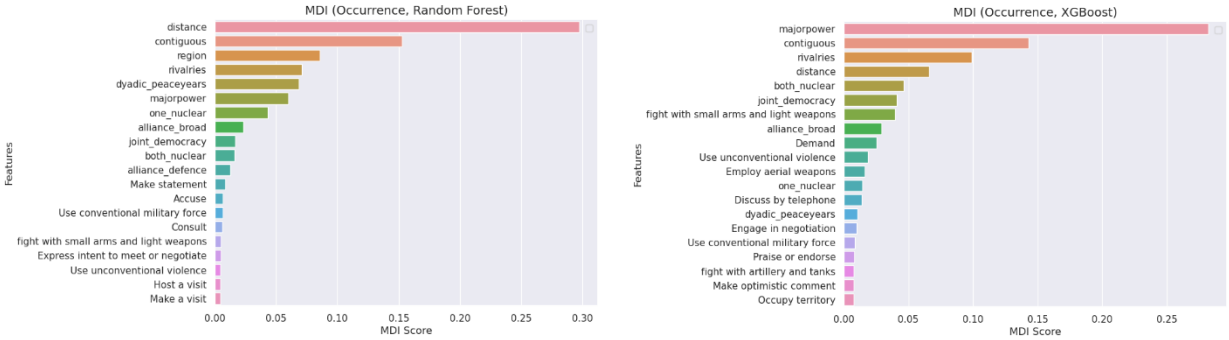


Figure 4-6. Feature importance RF and XGBoost – occurrence

Finding 2) is further corroborated when the feature rankings are summed across both algorithms. The highest-ranking event count features do not come close to the highest-ranking structural features. This supports the analysis that ICEWS event counts do not contribute significantly to the performance of conflict occurrence prediction models.

Table 4-7. A rough measure of averaged feature importance – occurrence

Most important structural features	*Feature*	*Sum of ranking across both tables*
	Contiguous	4
	Distance	5
	Rivalries	7
	Major power	7
Most important event features	Feature	Sum of ranking
	Fight with small arms or light weapons	23
	Use unconventional violence	28
	Use conventional military force	30

[14] See, for example, Maoz and Russet (1993).

4.2.2 Onset

When interpreting the feature importance for onset, we must keep in mind that the onset models had little predictive power in general. This inhibits drawing conclusions from the feature importances. Breiman, the inventor of the MDI score used here, said in his Wald Speech: "The better the model fits the data, the more sound the inferences about the black box are" (Breiman, 2002, p. 4). Conversely in our case, the worse the model fits the data, the less sound the inferences about the black box are. Therefore, the combined rankings of both algorithms MDI scores are not calculated for conflict onset, so as not to draw the wrong conclusions.

That being said, it is notable that random forest hinges heavily on the numeric features, while XGBoost includes contiguity as its main feature. The random forest behaviour likely reflects the bias of the MDI scores towards numeric features, whereas the XGBoost behaviour presents some further support for the importance of contiguity as a predictor.

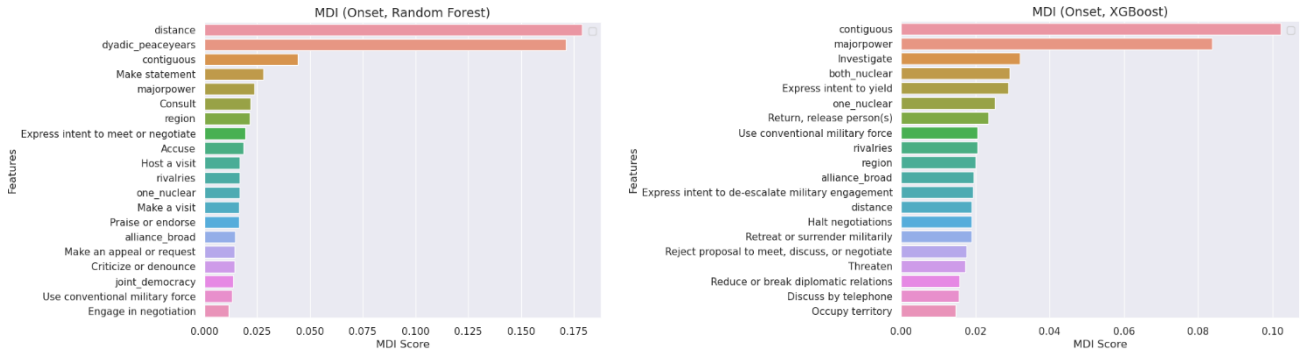


Figure 4-7. Feature importance RF and XGBoost – onset

4.2.3 Escalation

Similar to onset, distance is the most important variable in the random forest model, but is not used substantially by XGBoost. But contrary to onset, the escalation models do have predictive power. Once again, however, dyadic peace years is operationalised numerically, meaning that it offers many possibilities for splits, and it is favoured by the MDI procedure. It refers to the distance between the capitals or the nearest large city of the dyad members. Whether this variable is very important, is thus hard to conclude. What can be seen, however, is that contiguity, as in the MDI scores for the occurrence problem, has an important role in fitting the model to the training data.

As for the event counts, they are used slightly more than in occurrence, but by and large still do not form any substantive part of the predictive power of the models, further indicating that ICEWS event count features do not contribute to structural predictive models, as can be seen in the fact that their MDI rankings are substantially lower than the best-performing structural features.

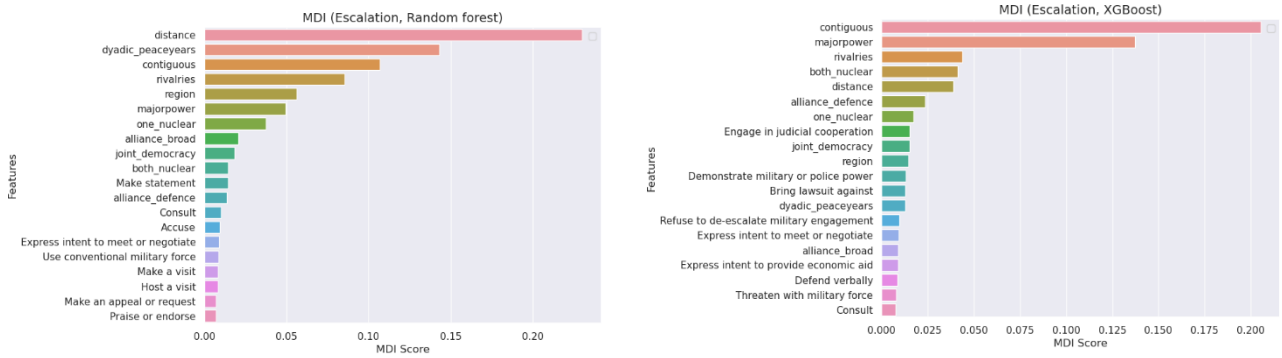


Figure 4-8. Feature importance RF and XGBoost - escalation

Table 4-8. A rough measure of averaged feature importance – escalation

Top-ranking features	*Feature*	*Sum of ranking across both tables*
	Contiguous	4
	Distance	6
	Rivalries	7
	Major power	8
Most important event features	Feature	Sum of ranking
	Express intent to meet or negotiate	31
	Consult	33

4.3 Model validity

4.3.1 Sensitivity train/test split

The sensitivity to the train/test split was tested by making three separate splits and running the random forest algorithm for conflict occurrence with all features. The main result is that, although there are some small differences in the different runs, the maximal difference is a 0.010 score. This indicates that the evaluation results in this report, although all based on the default split, do not depend heavily on the train/test split made. Nevertheless, it must be kept in mind that other evaluation outcomes inherently possess some variance, which is not displayed in the results reported.

Table 4-9 Results sensitivity train/test split (occurrence using random forest)

	Default split	*Alternative 1*	*Alternative 2*
Tuned hyperparameters			
n estimators	200	2000	1500
min. samples split	2	2	3
min. impurity decrease	0.0	0.0	0.0
max. features	square root	square root	square root
max. depth	50	55	50
criterion	gini	entropy	entropy
bootstrap	true	true	true
In-sample			
F2	1.000	0.999	0.998
AP	1.000	1.000	1.000
ROC AUC	1.000	1.000	1.000
MCC	1.000	0.999	0.997
Accuracy	1.000	0.999	0.998
Out-of-sample			

F2	0.970	0.964	0.967
AP	0.988	0.990	0.990
ROC AUC	0.986	0.988	0.990
MCC	0.934	0.924	0.925
Accuracy	0.967	0.962	0.962

* 5-fold cross-validation, 100 iterations

4.3.2 Overfitting and other validation measures

The models were checked for overfitting by comparing the results on the training set with the results on the test set. Of the three, only onset showed serious overfitting. In-sample, the onset results were far better than out-of-sample, indicating severe overfitting. To a degree, this might mean that the data is not informative enough to fit a predictor. However, it would be better to use nested cross-validation, instead of the train-test split approach applied here, prior to drawing that conclusion. The other two problems, occurrence and escalation, show a minor degree of overfitting. This is not problematic, since it is a general phenomenon that the model fits better to the training data, which it has used for learning, than on the unseen test data.

Finally, the models were run with a different size of the training and test set. When the training sample comprised 70% of the original data, versus the 67% used in the study, no effect on model performance was seen. This indicates that the 67% used as training sample is a reasonable choice of sample size.

5 DISCUSSION

5.1 Key findings and interpretation

The main question this study has tried to answer is "How can machine learning techniques and automated event data be employed to better predict and understand the onset and escalation of militarized interstate disputes?". For the occurrence of interstate conflict, it is clear that it is possible to predict it on a monthly level, instead of the yearly level that is common, thus far. It is also clear that even on a more disaggregated timescale, it is possible to achieve better performance then has been achieved until now. The prediction of onset of interstate conflict, however, remains very much troublesome. Even though there is space for improvement of the models, it is not unthinkable that this phenomenon cannot be accurately predicted in the shape of a classification problem. What is interesting however, is that the best occurrence model has only 46 cases of missed conflict, out of around 1500 or so. The best onset model misses 238 of the 268 conflict onsets. This means that the occurrence models are better at predicting conflict onset than the onset models. If this finding remains robust, a workaround for onset early-warning systems may simply be predicting occurrence cases. Finally, for escalation, we have seen that these methods are reasonably able to distinguish between the various levels of escalation of interstate conflict.

Subsequently, as for the effect of including automated event data in the input data. Can we 'catch' the short-term triggers that set off conflict when the conditions are right? Unfortunately, the effect of including the event counts had a negligible, or slightly negative, effect on the model performances across the prediction of occurrence, onset, and escalation. The possible reasons for this vary from insufficient testing of different operationalisations to too much noise in the feature space. More attention is paid to resolve these issues in subsection 5.4.1.

Finally, the feature importance scores indicate that indeed the structural variables do the heavy lifting, even when paired with AED. Within these variables, the MDI scores do not allow very specific conclusions about the influence of features, but nevertheless, contiguity is confirmed to be an effective predictor of interstate conflict. It ranks high in the MDI scores for both conflict occurrence and conflict escalation, the two problems with sufficiently powerful models to base conclusions on.

5.1.1 Comparison with other studies

The model of occurrence presents a slight upgrade to the state-of-the-art. Stodola et al. (2021) also used CoW data to predict MIDs with structural variables, and with the random forest algorithm. Contrary to the approach of this project however, they constructed a dataset out of politically relevant dyads, a scope that is somewhat narrower, which reduces the amount of noise in the dataset. Additionally, their unit of analysis was the dyad year, not the dyad month. Finally, the authors did not use the F_{beta} -score, nor average precision, so to compare the studies the ROC AUC is the best metric, keeping in mind that this metric does not do justice to the fact that false negatives carry more consequence than false positives.

As Table 5-1 shows, the current models constitute a minor upgrade to the latest models.

Table 5-1. Comparison earlier research – occurrence - ROC AUC

Train – test sample	*1:1*	*1:2*	*1:3*
Random forest (all features)	0.976	-	-
Stodola et al. (2021)	0.939	0.947	0.945

5.2 Implications for the scientific community and the policy world

The main implication of these findings for a policy application is that the models for conflict onset cannot be used as the basis of an early-warning system. It misses far too many conflict cases, which defeats its purpose. Additionally, pushing such a model into production may harm the reputation of conflict early-warning systems, given that policy makers may be alienated by data-heavy approaches even with functioning models (Sweijs & Teer, 2022).

Secondly, the models for occurrence might be used as an early-warning system. Their performance is good enough. However, two drawbacks still exist. First, occurrence is a relatively uninformative concept. When conflict has erupted on a certain dyad, no policy maker will be impressed by the prediction that there will be conflict next month as well. Second, these models do not identify the drivers of conflict sufficiently well, since the MDI scores do not provide such information. This does not allow policy makers access to levers for early action. Nevertheless, it has been shown that predicting occurrence on a monthly level is possible. If this model can be combined with more reliable feature importance scores, it might provide a workaround for the prediction of onset, with levers for early action. The application of the escalation models follows the same line as the occurrence models. The concept is not informative enough to base an early-warning system on, since escalation refers to the highest level of escalation experienced during an MID, which is not necessarily the level of escalation in the dyad month at hand.

The main finding of relevance for to the academic world is that, counterintuitively, the addition of AED did not improve any of the predictive models. However, it would be too soon to argue that AED has no role to play in predicting interstate conflict, as there is much room for improvement in this regard, see subsection 5.4.1. Moving on, this study does show that there is potential to predict interstate conflict at a more fine-grained level of temporal aggregation than was previously the case. This holds for both occurrence and escalation prediction. Additionally, it seems likely that it is possible to predict intraconflict dynamics with the right intraconflict data, when with a cruder measure of escalation cases of interstate war could be fairly well separated from other levels of escalation.

Finally, this study takes another step into finding out the limits of conflict prediction, as formulated by Chadefaux (2017). Whether predicting interstate conflict onset and escalation is (im)possible, is too soon to say, but at least there are indications that further improvements can be made. It seems that in that regard, we are one step closer to modelling elusive conflict, indicating that they behave as unpredictable clouds, about which we can at least assess some information about their probability distributions, rather than black swans. Of course, this study considered a time period without major changes to the international system (1995 – 2014), and all that would be necessary to invalidate any such hopes would be a black swan event.

5.3 Limitations

The study is subject to a number of limitations, some stemming from the data setup, others from the methods used. These are outlined below.

First, the model is not tested extensively across all available data. Instead of using different data samples from the original data, all analyses were run on a single under sampled dataset of around 10,000 dyad months, while the time period under study comprises more than 4 million. Second, within this sample, all models were then tested using a

single train/test split, instead of a nested cross-validation procedure. However, the sensitivity to the split was tested in subsection 4.3.1.

Then, one of the key aims of the study was to provide more insight into what factors are important for conflict prediction. However, the methodological setup of this study limits the interpretation of feature scores to indirect effects. This is not enough to decide on the effects of real-world factors. But there are better ways of looking into a black box. One that is promising is conditional permutation importance (Debeer & Strobl, 2020), with which the partial effects of variables can be calculated, i.e. the procedure can deal with multicollinearity. This may then be extended with partial dependence plots, with which one can assess the direction of a relationship, something which is impossible with feature importances. Alternatively, one can remove the multicollinearity by removing selected features and using regular permutation importance (Breiman, 2001). To remove multicollinearity, multiple methods exist, such as recursive correlation pruning or clustering features and selecting only one feature from each cluster. However, whatever procedure is applied, every explanation of model behaviour constructed from outside the model must be inherently wrong, in some aspects at least (Rudin, 2019). Otherwise, it would be an exact copy of the original – black box – model.

Subsequently, an aspect that plagues every data-intensive project, is that data are an imperfect representation of the real world, and even more so is the model built on it. Within this project, some aspects of the real world are crucially not incorporated. In part, this is due to the methods used, but nevertheless, there are components that heavily influence conflict behaviour that were not taken in. First, the dyad month as the unit of analysis precluded accounting for time structure and the presence of conflicts larger than a single dyad. While naturally, countries and country pairs are dependent on time, this aspect was removed from the unit of analysis, thus losing the capability to track conflict on a dyad through multiple dyad months. Normally, conflicts do not take place in isolated dyad months. The same goes for the spatial aspect. The dyad month format is unable to discern a world war from an isolated war, for instance. Subsequently, the data format does not account for a link between conflict onset and escalation and treats them as separable. Of course, an escalation in a relationship or dispute depends heavily on a previous onset of conflict. This effect was accounted for only indirectly, by measuring the number of dyadic peace years. Finally, the time period of the data is limited to 1995-2014. This means that the model has not been able to learn relations between predictors and conflict in the years 2015-2023. It is therefore reasonable to assume that the model would perform worse on real-time prediction than on the historical data.

5.4 Recommendations

5.4.1 Implementation and model extension

This study has made a first effort at incorporating automated event data into predictive models for interstate conflict occurrence, onset, and escalation. With the current scores, the model cannot be used for prediction of onset or intraconflict escalation for an early-warning system. However, there is reason to believe that the onset models can be sufficiently improved. The same does not go for the escalation models, since the manner in which the escalation outcome is measured does not allow intraconflict prediction. It is uncertain whether the onset models will be useful after including the following improvements. Nevertheless, they are the best options for implementing the current models.

1. **Extend AED operationalisation and use feature selection in onset models.**

The largest share of potential improvements is with the onset models. To start, the event counts performance may be improved by any combination of the following measures. First, the event counts might be adapted to account for the changing volume of recorded events throughout ICEWS history. ICEWS coverage has differed significantly due to expansion of covered news sources, or due to budget cuts, reducing the data volume. The event counts currently are distorted by this ever-changing level of coverage. Second, different aggregations of AED might be tried to increase performance, such as the difference between an event count and a rolling mean of the last months of event counts. This might better capture abrupt short-term changes. Third, the aggregate Goldstein scores a dyad's interactions might be included as a predictor. As an individual predictor, it does improve upon chance, although this is not presented in this report, so it might add to the AED's power. Third, a subset of relevant features can be extracted

before modelling, which reduces the noise in the feature space, and might improve prediction. This is a somewhat recent development in machine learning and is gaining popularity (Kamiri & Mariga, 2021). When combined with measures to remove multicollinearity, this will also improve the interpretability of features importances, especially when permutation feature importances are used (Tolоşi & Lengauer, 2011).

Finally, it might be possible to increase general performance in the binary problems by optimising the classification threshold, much like a hyperparameter. Classification models assign predicted classes to data samples, but this is based on an underlying probability. In this study, a probability threshold of 0.5 was used, meaning that new data points that were assigned a probability higher than 0.5, were marked as conflict cases., and vice versa. However, this split, while intuitive, might not be the best split. Perhaps the models can predict more samples correctly if the threshold were to be 0.3, or 0.55. This can be evaluated by tuning the threshold in cross-validation[15]. Especially for conflict onset, which suffers from class imbalance, this might improve performance. The improvements, however, should be expected to be incremental rather than dramatic, since the overall data structure does not change and the model learning remains similar.

2. Await update CoW MID data

The second recommendation is passive in nature. The Dyadic MID data from the CoW project were last updated in 2021 (Maoz et al., 2019). An update dataset closer to the current time might allow the occurrence models to be used in real-time to predict next month's conflict occurrence cases.

3. Implement cost-sensitive training

This extension might offer substantial improvements without the need for changing the data setup, as in final and last recommendation. In cost-sensitive training, the loss function used by the algorithm differentiates between different errors, i.e., it assigns different costs to misclassifying different classes. In our case, the loss function should assign high costs to false negatives, conflicts that did occur, but were predicted as peace. Cost-sensitive training is different from the approach used in this study. Here, the model parameters were optimised for predicting negative classes by using the F2-score as the performance metric in cross-validation, but the training phase itself did not differentiate between false positives or false negatives. By altering the loss function, cost-sensitive training has the potential to change, and perhaps improve, the classifier substantially (Kuhn & Johnson, 2013, pp. 429–435).

4. Extend test set to real-world ratio of peace and conflict

To increase model validity the trained models should be tested on the full sample of available dyad months. This would expose the model to the real-world ratio of conflict and peace and would allow for a much more robust evaluation. The most effort would be required to rewrite the data manipulation code for the structural variables to allow out-of-memory computation, and to adapt the evaluation metrics to account for severe class-imbalance in the test set.

5.4.2 Recommendations for future research

1. Use Ensemble Bayesian Model Averaging for interstate conflict onset prediction

Ensemble Bayesian Model Averaging contains two key aspects. It is a model ensemble, and it is a Bayesian technique. Model ensembles have been gaining traction in the conflict prediction literature with successful implementations in the state-of-the-art ViEWS project (Hegre, Bell, et al., 2021), and in predicting conflict intensity in Africa (Ettensperger, 2021). Model ensembles have the advantage that they can combine multiple models, each with their own strengths, although this comes at a cost of interpretability. Bayesian methods have been shown to be

[15] See Williford and Atkinson (2019) for an explanation of the importance of choosing a reasonable threshold in conflict prediction. They do not optimise the threshold value, but they do showcase its effect on a model's behaviour, using the ROC AUC.

successful as well. They have the advantage that they can include expert knowledge in the form of informative priors to reduce noise in the model. Williford and Atkinson (2019) used Bayesian model averaging to predict interstate conflict onsets, with fairly good results.

2. Host a forecasting competition for interstate conflict onset

The number of studies forecasting explicitly interstate conflict onset is still fairly low compared to other forms of conflict forecasting. A particular recommendation for the Hague Centre of Strategic Studies, or any other policy modelling institute, is hosting a forecasting competition for the prediction of interstate conflict onset. The ViEWS project has hosted a competition for predicting escalation of state-based violence in Africa (Vesco et al., 2022). They found that when combining the heterogeneous models of the participating teams, the resulting ensemble model performed better than each individual model. Additionally, such a focused effort on a single problem results in many opportunities for furthering the field.

6 CONCLUSIONS

The goal of this study was to improve predictive models for interstate conflict. To that end, it tested whether the addition of automated event data improved upon the current models which rely on structural variables. For all three problems, conflict occurrence, conflict onset, and conflict escalation, there was no benefit found to including automated event data. Nevertheless, it cannot be stated that automated event data have no use in interstate conflict prediction within this setup, because further tests will need to be carried out to confirm this.

The research process started with selecting structural predictors on the basis of theorised causal effects to interstate conflict. This selection criterion is suitable, since the structural predictors are capable of predicting the occurrence and escalation of conflict with reasonable accuracy. The predictor with the most robust effect on all problems is contiguity. The approach used was a machine learning approach with tree ensembles, selected for their flexibility and a certain measure of interpretability.

The occurrence and escalation models were found to be performing somewhat above the benchmark in the field, but the models for conflict occurrence do not perform well. Based on the outcomes of this study, the recommendations for further research are 1) to continue testing automated event data setups, 2) to use model ensemble , more specifically Bayesian ensembles, given their performance in the field, and 3) to host a prediction competition for interstate conflict at the Hague Centre of Strategic Studies, or any other policy modelling institution.

The main contribution of this study is that automated event data combined with machine learning and structural data do not improve interstate conflict prediction. However, there is potential in predicting escalation levels throughout an interstate conflict, and it also has been shown that it is possible to model interstate conflict on a monthly level instead of a yearly level.

Early-warning systems accurately predicting conflict onset and its current escalation with a month lead time, linked to actionable policy levers, remain a goal for the time being. Nevertheless, accurate prediction of conflict occurrence, possibly linked to early action levers are within reach. Finally, predicting escalation levels does seem to be possible as well.

7 REFERENCES

LEETARU v. THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ILLINOIS, AND GUENTHER, No. 4-13-0290 (4th. District, Appellate Court, Illinois 29 April 2013). https://www.illinoiscourts.gov/Resources/5ee30570-9d30-48a4-ad11-eb38060f04f8/4130290_R23.pdf

Achen, C. H. (2006). Evaluating political decision-making models. In C. H. Achen, F. N. Stokman, R. Thomson, & T. König (Eds.), *The European Union Decides* (pp. 264–298). Cambridge University Press. https://doi.org/10.1017/CBO9780511492082.011

Arva, B., Beieler, J., Fisher, B., Lara, G., Schrodt, P. A., Song, W., Sowell, M., & Stehle, S. (2013, July 3). *Improving Forecasts of International Events of Interest*. European Political Studies Association meetings. https://parusanalytics.com/eventdata/papers.dir/Arva.etal_EPSA_13.pdf

Azar, E. E. (1980). The Conflict and Peace Data Bank (COPDAB) Project. *Journal of Conflict Resolution*, *24*(1), 143–152. https://doi.org/10.1177/002200278002400106

Azar, E. E., McLaurin, R. D., Havener, T., Murphy, C., Sloan, T., & Wagner, C. H. (1977). A System for Forecasting Strategic Crises: Findings and Speculations About Conflict in the Middle East. *International Interactions*, *3*(3), 193–222. https://doi.org/10.1080/03050627708434464

Beck, N., King, G., & Zeng, L. (2000). Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review*, *94*(1), 21. https://doi.org/10.2307/2586378

Bennett, D. S., & Stam, A. C. (2000). Eugene: A conceptual manual. *International Interactions*, *26*(2), 179–204. https://doi.org/10.1080/03050620008434965

Bennett, D. S., & Stam, A. C. (2004). *The Behavioral Origins of War*. University of Michigan Press. https://doi.org/10.3998/mpub.23490

Blair, R. A., Blattman, C., & Hartman, A. (2017). Predicting local violence: Evidence from a panel survey in Liberia. *Journal of Peace Research*, *54*(2), 298–312. https://doi.org/10.1177/0022343316684009

Bond, D., Bond, J., Oh, C., Jenkins, J. C., & Taylor, C. L. (2003). Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development. *Journal of Peace Research*, *40*(6), 733–745. https://doi.org/10.1177/00223433030406009

Boulesteix, A.-L., Bender, A., Lorenzo Bermejo, J., & Strobl, C. (2012). Random forest Gini importance favours SNPs with large minor allele frequency: Impact, sources and recommendations. *Briefings in Bioinformatics*, *13*(3), 292–304. https://doi.org/10.1093/bib/bbr053

Braithwaite, A., & Lemke, D. (2011). Unpacking Escalation. *Conflict Management and Peace Science*, *28*(2), 111–123. https://doi.org/10.1177/0738894210396631

Brandt, P. T., D'Orazio, V., Khan, L., Li, Y.-F., Osorio, J., & Sianan, M. (2022). Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions*, *48*(4), 800–822. https://doi.org/10.1080/03050629.2022.2036987

Brandt, P. T., & Freeman, J. R. (2006). Advances in Bayesian Time Series Modeling and the Study of Politics: Theory Testing, Forecasting, and Policy Analysis. *Political Analysis*, *14*(1), 1–36. https://doi.org/10.1093/pan/mpi035

Brecher, M., & Wilkenfeld, J. (2000). *A Study of Crisis*. University of Michigan Press. https://doi.org/10.3998/mpub.14982

Brecher, M., Wilkenfeld, J., Beardsley, K., James, P., & Quinn, D. M. (2021). *International Crisis Behavior Data Codebook, Version 14*. Duke University. https://sites.duke.edu/icbdata/data-collections/

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2002). *Wald Lecture II Looking Inside the Black Box*. https://www.stat.berkeley.edu/~breiman/wald2002-2.pdf

Breiman, L. (2003). *Manual On Setting Up, Using, And Understanding Random Forests V3.1*. https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf

Bremer, S. A. (1992). Dangerous Dyads: Conditions Affecting the Likelihood of Interstate War, 1816-1965. *Journal of Conflict Resolution*, *36*(2), 309–341. https://doi.org/10.1177/0022002792036002005

Cederman, L.-E., & Gleditsch, K. S. (2009). Introduction to Special Issue on "Disaggregating Civil War". *Journal of Conflict Resolution*, *53*(4), 487–495. https://doi.org/10.1177/0022002709336454

Cederman, L.-E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, *355*(6324), 474–476. https://doi.org/10.1126/science.aal4483

Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, *51*(1), 5–18. https://doi.org/10.1177/0022343313507302

Chadefaux, T. (2017a). Market anticipations of conflict onsets. *Journal of Peace Research*, *54*(2), 313–327. http://www.jstor.org/stable/44511213

Chadefaux, T. (2017b). Conflict forecasting and its limits. *Data Science*, *1*(1–2), 7–17. https://doi.org/10.3233/DS-170002

Chadefaux, T. (2022). A shape-based approach to conflict forecasting. *International Interactions*, *48*(4), 633–648. https://doi.org/10.1080/03050629.2022.2009821

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chiba, D., & Gleditsch, K. S. (2017). The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research*, *54*(2), 275–297. http://www.jstor.org/stable/44511211

Choucri, N. (1974). Forecasting in international relations: Problems and prospects. *International Interactions*, *1*(2), 63–86. https://doi.org/10.1080/03050627408434390

Choucri, N., & Robinson, T. W. (1978). *Forecasting in International Relations: Theory, Methods, Problems, Prospects*. W. H. Freeman.

Collier, P., & Hoeffler, A. (2004). Greed and Grievance in Civil War. *Oxford Economic Papers*, *56*(4), 563–595. http://www.jstor.org/stable/3488799

Correlates of War Project. (n.d.). *Direct Contiguity Data, 1816—2016. Version 3.2*. Correlates of War. Retrieved 7 July 2022, from https://correlatesofwar.org/data-sets/direct-contiguity

Correlates of War Project. (2017). *State System Membership List, v2016*. Correlates of War Project. https://correlatesofwar.org/data-sets/state-system-membership/

Correlates of War Project. (2017). *State System Membership List, v2016* [Folder]. Correlates of War. http://correlatesofwar.org

Daxecker, U., & Prins, B. C. (2017). Financing rebellion: Using piracy to explain and predict conflict intensity in Africa and Southeast Asia. *Journal of Peace Research*, *54*(2), 215–230. https://www.jstor.org/stable/44511207

Debeer, D., & Strobl, C. (2020). Conditional permutation importance revisited. *BMC Bioinformatics*, *21*(1), 307. https://doi.org/10.1186/s12859-020-03622-2

Diehl, P. F., & Goertz, G. (2000). *War and Peace in International Rivalry*. The University of Michigan Press. https://doi.org/10.3998/mpub.16693

Diehl, P. F., Goertz, G., & Gallegos, Y. (2021). Peace data: Concept, measurement, patterns, and research agenda. *Conflict Management and Peace Science*, *38*(5), 605–624. https://doi.org/10.1177/0738894219870288

Douglass, R. W., Scherer, T. L., Gannon, J. A., Gartzke, E., Lindsay, J., Carcelli, S., Wilkenfeld, J., Quinn, D. M., Aiken, C., Navarro, J. M. C., Lund, N., Murauskaite, E., & Partridge, D. (2022). *Introducing the ICBe Dataset: Very High Recall and Precision Event Extraction from Narratives about International Crises*. https://doi.org/10.48550/arXiv.2202.07081

Ettensperger, F. (2021). Forecasting conflict using a diverse machine-learning ensemble: Ensemble averaging with multiple tree-based algorithms and variance promoting data configurations. *International Interactions*, 1–23. https://doi.org/10.1080/03050629.2022.1993209

EU. (2020). *Factsheet: EU conflict Early Warning System*. European Union. https://www.eeas.europa.eu/sites/default/files/ews_fact_sheet_2020.pdf

Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, Insurgency, and Civil War. *American Political Science Review*, *97*(1), 75–90. https://www.jstor.org/stable/3118222

Frederick, B. A., Hensel, P. R., & Macaulay, C. (2017). The Issue Correlates of War Territorial Claims Data, 1816–20011. *Journal of Peace Research*, *54*(1), 99–108. https://doi.org/10.1177/0022343316676311

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference, 1996.*, 148–156. https://www.semanticscholar.org/paper/Experiments-with-a-New-Boosting-Algorithm-Freund-Schapire/68c1bfe375dde46777fe1ac8f3636fb651e3f0f8

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Gartzke, E. (2007). The Capitalist Peace. *American Journal of Political Science*, *51*(1), 166–191. https://doi.org/10.1111/j.1540-5907.2007.00244.x

Gerner, D. J., Schrodt, P. A., Francisco, R. A., & Weddle, J. L. (1994). Machine Coding of Event Data Using Regional and International Sources. *International Studies Quarterly*, *38*(1), 91–119. https://doi.org/10.2307/2600873

Gerner, D. J., Schrodt, P. A., Yilmaz, O., & Abu-Jabr, R. (2002). *The Creation of CAMEO (Conflict and Mediation Event Observations): An Event Data Framework for a Post Cold War World*. 32.

Gibler, D. M. (2009). *International Military Alliances, 1648-2008*. CQ Press. http://us.sagepub.com/en-us/nam/international-military-alliances-1648-2008/book236420

Gleditsch, K. S. (2017). Ornithology and Varieties of Conflict: A Personal Retrospective on Conflict Forecasting. *Peace Economics, Peace Science and Public Policy*, *23*(4). https://doi.org/10.1515/peps-2017-0023

Gleditsch, K. S., & Ward, M. D. (2013). Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, *50*(1), 17–31. https://doi.org/10.1177/0022343312449033

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., & Strand, H. (2002). Armed Conflict 1946-2001: A New Dataset. *Journal of Peace Research*, *39*(5), 615–637. https://doi.org/10.1177/0022343302039005007

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J., & Woodward, M. (2010). A Global Model for Forecasting Political Instability. *American Journal of Political Science*, *54*(1), 190–208. https://doi.org/10.1111/j.1540-5907.2009.00426.x

Halkia, M., Ferri, S., Schellens, M. K., Papazoglou, M., & Thomakos, D. (2020). The Global Conflict Risk Index: A quantitative tool for policy support on conflict prevention. *Progress in Disaster Science*, *6*, 100069. https://doi.org/10.1016/j.pdisas.2020.100069

Harff, B. (2003). No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder since 1955. *American Political Science Review*, *97*(1), 57–73. https://www.jstor.org/stable/3118221

Hegre, H. (2008). Gravitating toward War: Preponderance May Pacify, but Power Kills. *Journal of Conflict Resolution*, *52*(4), 566–589. https://doi.org/10.1177/0022002708316738

Hegre, H. (2014). Democracy and armed conflict. *Journal of Peace Research*, *51*(2), 159–172. https://doi.org/10.1177/0022343313512852

Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., Hoyles, F., Hultman, L., Högbladh, S., Jansen, R., Mouhleb, N., Muhammad, S. A., Nilsson, D., Nygård, H. M., Olafsdottir, G., Petrova, K., Randahl, D., Rød, E. G., Schneider, G., … Vestby, J. (2019). ViEWS: A political violence early-warning system. *Journal of Peace Research*, *56*(2), 155–174. https://doi.org/10.1177/0022343319823860

Hegre, H., Bell, C., Colaresi, M., Croicu, M., Hoyles, F., Jansen, R., Leis, M. R., Lindqvist-McGowan, A., Randahl, D., Rød, E. G., & Vesco, P. (2021). ViEWS2020: Revising and evaluating the ViEWS political Violence Early-Warning System. *Journal of Peace Research*, *58*(3), 599–611. https://doi.org/10.1177/0022343320962157

Hegre, H., Bernhard, M., & Teorell, J. (2020). Civil Society and the Democratic Peace. *Journal of Conflict Resolution*, *64*(1), 32–62. https://doi.org/10.1177/0022002719850620

Hegre, H., Metternich, N. W., Nygård, H. M., & Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research*, *54*(2), 113–124. https://doi.org/10.1177/0022343317691330

Hegre, H., Nygård, H. M., & Landsverk, P. (2021). Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand Up to Reality. *International Studies Quarterly*, *65*(3), 660–668. https://doi.org/10.1093/isq/sqaa094

Heidelberg Institute International Conflict Relations Research. (2021). *Conflict Barometer 2020*. Universität Heidelberg. https://hiik.de/conflict-barometer/bisherige-ausgaben/?lang=en

Henrickson, P. (2020). Predicting the costs of war. *The Journal of Defense Modeling and Simulation*, *17*(3), 285–308. https://doi.org/10.1177/1548512919826375

Hobson, C. (2017). Democratic Peace: Progress and Crisis. *Perspectives on Politics*, *15*(3), 697–710. https://doi.org/10.1017/S1537592717000913

Jones, D. M., Bremer, S. A., & Singer, J. D. (1996). Militarized Interstate Disputes, 1816–1992: Rationale, Coding Rules, and Empirical Patterns. *Conflict Management and Peace Science*, *15*(2), 163–213. https://doi.org/10.1177/073889429601500203

Kamiri, J., & Mariga, G. (2021). Research Methods in Machine Learning: A Content Analysis. *International Journal of Computer and Information Technology(2279-0764)*, *10*(2), Article 2. https://doi.org/10.24203/ijcit.v10i2.79

King, G., & Zeng, L. (2001). Explaining Rare Events in International Relations. *International Organization*, *55*(3), 693–715. https://doi.org/10.1162/00208180152507597

Kinsella, D. (2005). No Rest for the Democratic Peace. *American Political Science Review*, *99*(3), 453–457. https://doi.org/10.1017/S0003055405051774

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lemke, D., & Reed, W. (2001). The Relevance of Politically Relevant Dyads. *Journal of Conflict Resolution*, *45*(1), 126–144. http://www.jstor.org/stable/3176286

Maoz, Z. (1982). *Paths to Conflict: International Dispute Initiation, 1816-1976*. Westview Press.

Maoz, Z., Johnson, P. L., Kaplan, J., Ogunkoya, F., & Shreve, A. P. (2019). The Dyadic Militarized Interstate Disputes (MIDs) Dataset Version 3.0: Logic, Characteristics, and Comparisons to Alternative Datasets. *Journal of Conflict Resolution*, *63*(3), 811–835. https://doi.org/10.1177/0022002718784158

Maoz, Z., & Russett, B. (1993). Normative and Structural Causes of Democratic Peace, 1946-1986. *The American Political Science Review*, *87*(3), 624–638. https://doi.org/10.2307/2938740

Marshall, M. G., Jaggers, K., & Gurr, T. R. (2020). *The Polity Project*. Center for Systemic Peace. https://www.systemicpeace.org/polityproject.html

Marwala, T., & Lagazio, M. (2004). *Modeling and controlling interstate conflict*. *2*, 1233–1238 vol.2. https://doi.org/10.1109/IJCNN.2004.1380119

Marwala, T., & Lagazio, M. (2011). *Militarized Conflict Modeling Using Computational Intelligence* (1st ed.). Springer, London. https://link.springer.com/book/10.1007/978-0-85729-790-7

Mesquita, B. B. de. (1980). An Expected Utility Theory of International Conflict. *American Political Science Review*, *74*(4), 917–931. https://doi.org/10.2307/1954313

Mesquita, B. B. de. (1985). The War Trap Revisited: A Revised Expected Utility Model. *American Political Science Review*, *79*(1), 156–177. https://doi.org/10.2307/1956125

Meyer, C. O., De Franco, C., & Otto, F. (2019). *Warning about War: Conflict, Persuasion and Foreign Policy*. Cambridge University Press. https://doi.org/10.1017/9781108644006

Milstein, J. S. (1974). *Dynamics of the Vietnam war: A quantitative analysis and predictive computer simulation*. The Ohio State University Press. http://hdl.handle.net/1811/24664

Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving Predictions using Ensemble Bayesian Model Averaging. *Political Analysis*, *20*(3), 271–291. https://doi.org/10.1093/pan/mps002

O'Brien, S. P. (2010). Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, *12*(1), 87–104. http://www.jstor.org/stable/40730711

UN Office for the Coordination of Humanitarian Affairs. (2022). *Assessing the technical feasibility of conflict prediction for early action*. The Centre for Humanitarian Data. https://centre.humdata.org/assessing-the-technical-feasibility-of-conflict-prediction-for-anticipatory-action/

Organisation for Economic Co-operation and Development. (2009). *Preventing Violence, War and State Collapse: The Future of Conflict Early Warning and Response*. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264059818-en

Oneal, J. R., & Russett, B. M. (1997). The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950-1985. *International Studies Quarterly*, *41*(2), 267–293. http://www.jstor.org/stable/3013934

Palmer, G., McManus, R. W., D'Orazio, V., Kenwick, M. R., Karstens, M., Bloch, C., Dietrich, N., Kahn, K., Ritter, K., & Soules, M. J. (2022). The MID5 Dataset, 2011-2014: Procedures, Coding Rules, and Description. *Conflict Management and Peace Science*, *39*(4), 470–482. https://doi.org/10.1177/0738894221995743

Pettersson, T., Davies, S., Deniz, A., Engström, G., Hawach, N., Högbladh, S., & Öberg, M. S. M. (2021). Organized violence 1989–2020, with a special emphasis on Syria. *Journal of Peace Research*, *58*(4), 809–825. https://doi.org/10.1177/00223433211026126

Popper, K. R. (1979). *Objective Knowledge: An Evolutionary Approach*. Clarendon Press.

Reed, W. (2000). A Unified Statistical Model of Conflict Onset and Escalation. *American Journal of Political Science*, *44*(1), 84–93. https://doi.org/10.2307/2669294

Reiter, D., & Tillman, E. R. (2002). Public, Legislative, and Executive Constraints on the Democratic Initiation of Conflict. *The Journal of Politics*, *64*(3), 810–826. https://doi.org/10.1111/0022-3816.00147

Richardson, L. F. (1960). *Statistics of deadly quarrels* (2nd ed.). Boxwood Press. (Original work published 1949)

Rokach, L. (2009). *Pattern Classification Using Ensemble Methods* (Vol. 75). World Scientific. https://doi.org/10.1142/7238

Roser, M., Herre, B., & Hasell, J. (2013, August 6). *Nuclear Weapons*. Our World in Data. https://ourworldindata.org/nuclear-weapons

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Russett, B., Layne, C., Spiro, D. E., & Doyle, M. W. (1995). The Democratic Peace. *International Security*, *19*(4), 164–184. https://doi.org/10.2307/2539124

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, *10*(2), 153–178. https://doi.org/10.1007/BF00993504

Scharpf, A., Schneider, G., Nöh, A., & Clauset, A. (2014). Forecasting the Risk of Extreme Massacres in Syria. *European Review of International Studies*, *1*(2), 50–68. https://www.jstor.org/stable/26593335

Schneider, G., Barbieri, K., & Gleditsch, N. P. (2003). *Globalization and Armed Conflict*. Rowman & Littlefield. https://rowman.com/ISBN/9780742518322/Globalization-and-Armed-Conflict

Schneider, G., Gleditsch, N. P., & Carey, S. (2011). Forecasting in International Relations: One Quest, Three Approaches. *Conflict Management and Peace Science*, *28*(1), 5–14. https://doi.org/10.1177/0738894210388079

Schrodt, P. A. (1990). Predicting Interstate Conflict Outcomes Using a Bootstrapped ID3 Algorithm. *Political Analysis*, *2*, 31–56. https://doi.org/10.1093/pan/2.1.31

Schrodt, P. A. (1991). Prediction of Interstate Conflict Outcomes Using a Neural Network. *Social Science Computer Review*, *9*(3), 359–380. https://doi.org/10.1177/089443939100900302

Schrodt, P. A. (2001). *Automated Coding Of International Event Data Using Sparse Parsing Techniques*. International Studies Association, Chicago. https://www.researchgate.net/publication/2348151_Automated_Coding_Of_International_Event_Data_Using_Sparse_Parsing_Techniques

Schrodt, P. A. (2012). *CAMEO Conflict and Mediation Event Observations Event and Actor Codebook*. Parus Analytical Systems. https://parusanalytics.com/eventdata/data.dir/cameo.html

Schrodt, P. A. (2014). Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, *51*(2), 287–300. https://doi.org/10.1177/0022343313499597

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political Science: KEDS-A Program for the Machine Coding of Event Data. *Social Science Computer Review*, *12*(4), 561–587. https://doi.org/10.1177/089443939401200408

Schrodt, P. A., & Gerner, D. J. (2000). Cluster-Based Early Warning Indicators for Political Change in the Contemporary Levant. *American Political Science Review*, *94*(4), 803–817. https://doi.org/10.2307/2586209

Schrodt, P. A., & Yonamine, J. (2016). A Guide to Event Data: Past, Present, and Future. *All Azimuth: A Journal of Foreign Policy and Peace*, *2*(2), 5–5. https://doi.org/10.20991/allazimuth.167312

scikit-learn. (n.d.). *3.1. Cross-validation: Evaluating estimator performance*. Scikit-Learn 1.1.3 Documentation. Retrieved 23 November 2022, from https://scikit-learn.org/stable/modules/cross_validation.html

Singer, D. (1973). The peace researcher and foreign policy prediction. In *Advancing Peace Research* (1st ed., p. 11). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203128398-18/peace-researcher-foreign-policy-prediction-1973

Singer, J. D., & Wallace, M. D. (1979). *To augur well; early warning indicators in world politics*. Sage. https://unesdoc.unesco.org/ark:/48223/pf0000038285

Singh, S. (2018, May 21). *Understanding the Bias-Variance Tradeoff*. Medium. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Small, M., & Singer, J. D. (1982). *Resort to arms: International and civil wars, 1816-1980* ([2nd ed.]). Sage Publications. https://catdir.loc.gov/catdir/enhancements/fy0660/81018518-d.html

Stephens, J. (2012). Political Scientists Are Lousy Forecasters. *The New York Times*. https://www.nytimes.com/2012/06/24/opinion/sunday/political-scientists-are-lousy-forecasters.html

Stinnett, D. M., Tir, J., Diehl, P. F., Schafer, P., & Gochman, C. (2002). The Correlates of War (Cow) Project Direct Contiguity Data, Version 3.0. *Conflict Management and Peace Science*, *19*(2), 59–67. https://doi.org/10.1177/073889420201900203

Stodola, P., Vojtek, J., Kutěj, L., & Neubauer, J. (2021). Modelling militarized interstate disputes using data mining techniques: Prevention and prediction of conflicts. *The Journal of Defense Modeling and Simulation*, *18*(4), 469–483. https://doi.org/10.1177/1548512920925178

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Sundberg, R., & Melander, E. (2013). Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research*, *50*(4), 523–532. https://doi.org/10.1177/0022343313484347

Sweijs, T., & Teer, J. (2022). *Practices, Principles and Promises of Conflict Early Warning Systems* (p. 58). The Hague Centre for Strategic Studies. https://hcss.nl/report/practices-principles-and-promises-of-conflict-early-warning-systems/

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House Publishing Group.

Terechshenko, Z. (2020). Hot under the collar: A latent measure of interstate hostility. *Journal of Peace Research*, *57*(6), 764–776. https://doi.org/10.1177/0022343320962546

Tetlock, P. E. (2017). *Expert Political Judgment: How Good Is It? How Can We Know? - New Edition* (2nd ed.). Princeton University Press. https://press.princeton.edu/books/hardcover/9780691178288/expert-political-judgment (Original work published 2005)

Thompson, W. R. (2001). Identifying Rivals and Rivalries in World Politics. *International Studies Quarterly*, *45*(4), 557–586. http://www.jstor.org/stable/3096060

Thompson, W. R., Sakuwa, K., & Suhas, P. H. (2022). *Analyzing Strategic Rivalries in World Politics*. Springer Nature Singapore. https://link.springer.com/book/10.1007/978-981-16-6671-1

Tolоşi, L., & Lengauer, T. (2011). Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300

Ulfelder, J. (2014, March 20). On Prediction. *Dart-Throwing Chimp*. https://dartthrowingchimp.wordpress.com/2014/03/20/on-prediction/

UN. (1992). *An agenda for peace: Preventive diplomacy, peacemaking and peace-keeping*. UN Department of Public Information,. https://digitallibrary.un.org/record/145749

UN. (2011). *Preventive diplomacy: Delivering results*. United Nations. https://digitallibrary.un.org/record/710438

Vesco, P., Hegre, H., Colaresi, M., Jansen, R. B., Lo, A., Reisch, G., & Weidmann, N. B. (2022). United they stand: Findings from an escalation prediction competition. *International Interactions*, *48*(4), 860–896. https://doi.org/10.1080/03050629.2022.2029856

Ward, M. D. (2016). Can We Predict Politics? Toward What End? *Journal of Global Security Studies*, *1*(1), 80–91. https://doi.org/10.1093/jogss/ogv002

Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, *47*(4), 363–375. https://doi.org/10.1177/0022343309356491

Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., & Weschle, S. (2013). Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction. *International Studies Review*, *15*(4), 473–490. http://www.jstor.org/stable/24032984

Weidmann, N. B., & Ward, M. D. (2010). Predicting Conflict in Space and Time. *Journal of Conflict Resolution*, *54*(6), 883–901. https://doi.org/10.1177/0022002710371669

Williford, G. W., & Atkinson, D. B. (2019). A Bayesian forecasting model of international conflict. *The Journal of Defense Modeling and Simulation*, *17*(3), 235–242. https://doi.org/10.1177/1548512919827659

Witmer, F. D., Linke, A. M., O'Loughlin, J., Gettelman, A., & Laing, A. (2017). Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models. *Journal of Peace Research*, *54*(2), 175–192. https://doi.org/10.1177/0022343316682064

World Bank Group. (2018). *Atlas of Sustainable Development Goals 2018: From World Development Indicators*. World Bank Group. https://doi.org/10.1596/978-1-4648-1250-7

Wright, Q. (1965). *A study of war* (2nd ed.). University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/S/bo3612694.html (Original work published 1942)

XGBoost Documentation—Xgboost 1.7.2 documentation. (2022). https://xgboost.readthedocs.io/en/stable/

Appendix A DATA

A.1 Target variables

A.1.1 Selecting conflict dataset

The target variable data are sourced from the MID Dyadic dataset. It offers the best distribution of conflict cases across escalation levels, and has the most convenient format. In the sections below, the conditions used in selecting a dataset are set out, after which the candidate conflict datasets are each evaluated against those conditions in the next section

A.1.1.1 Conditions for selection

There are a number of conditions that apply in selecting the dataset to be used. Three of those have a hard boundary, the dataset either meets or does not meet the condition. One condition has a soft boundary, meaning that the degree to which the condition is met is the deciding factor. All four, however, are critical to the project's successful execution.

The first of these conditions is that 1) the data must contain cases explicitly coded as interstate conflict. There is a blossoming literature on the prediction of various forms of conflict, such as civil wars, ethnic violence, and intrastate violence (see for example the ViEWS project, which includes state-of-the-art models for predicting intrastate violence (Hegre, Bell, et al., 2021), and the Global Conflict Risk Index, the EU early-warning model for intrastate violence (Halkia et al., 2020). However, this means that not every conflict dataset explicitly measures the interstate conflict required for this research. Then, 2), the data must allow for disaggregation to the level of the dyad month. This is the unit of analysis of the project. A start date and an end date for any conflict must be included and must be accurate to at least the level of a month. Subsequently, 3) the data must allow for constructing a measure of escalation. In the wider literature, there is no consensus on a measure for the concept of escalation of interstate conflict. Different studies use different measures, often using the number of fatalities, use of force, or the reciprocity of a conflict (Braithwaite & Lemke, 2011; Terechshenko, 2020). Therefore, our datasets are judged as to what extent they offer information on these conflict characteristics. The exact measure used depends on what is available for our purposes, and on the distribution of the data points through various classes of escalation, as mentioned in condition four. The choice for final measure is explained in section 3.1.2.1. The last-to-one condition 4) pertains to the size of the available data. The data must have as many usable data points as possible, when transformed into a dyad month format. The minimum is a couple of hundred cases, available from 1995 onwards, since this is the starting point of the ICEWS project data collection, the event dataset that is used. This condition also applies to the different classes of the escalation scale to be constructed. A dataset that offers a decent distribution across different classes of escalation offers greater flexibility in successfully predicting various levels of escalation. And finally, 5) the data's scope must be global, since the model aims for global prediction, and extend in time as far as possible from 1995 onwards.

A.1.1.2 Overview of conflict datasets

Given the exact requirements and conditions outlined in the previous section, the MID Dyadic Dataset Version 4.02 is most suited to the research objectives. The following section presents an overview of the candidate conflict datasets, and the arguments pro and contra using them. These eight datasets were selected after a background study, and have all been constructed manually. These are in general more accurate compared to their automated counterpart, because humans are better than machines at taking the context of historical events into account. The classifying power of the models depends in large part on the target variable's accuracy, hence this choice for manually constructed datasets.

CoW Militarized Interstate Dispute Data, version 5.0 (MIDA)

The Correlates of War project's MID dataset (Palmer et al., 2022) was first published in 1984, and receiving multiple updates through the years, has become the standard for quantitative research into interstate conflict. It offers data from 1816 until 2014 on MIDs, which are defined as "united historical cases of conflict in which the threat, display or use of military force short of war by one member state is explicitly directed towards the government, official representatives, official forces, property, or territory of another state" (Jones et al., 1996, p. 163). It measures the number of fatalities, whether force was used, and whether a conflict was reciprocated. Additionally, it offers the starting and ending dates of a conflict.

CoW MID Incidents Dataset, version 5.0 (MIDI)

The MIDI dataset contains the incidents that make up the MIDs described in the MIDA dataset, but only for the time period 1993-2014. Incidents are the building blocks of MIDs, and they are defined as a "single military action involving an explicit threat, display, or use of force by one system member state towards another system member state" (Jones et al., 1996, p. 169). Incident-level data allows prediction of escalation levels within a conflict, similar to the ICBe data later to be discussed. Nevertheless, as Table A-1 shows, MIDI also has a significant shortcoming in the fact that the highest escalation class has a very low event count, making effective prediction of escalation to war impossible. The incident level data does offer an opportunity for future research, focussing on prediction of escalation through time.

MID Dyadic Dataset, version 4.02

The MID Dyadic dataset is built almost entirely on the MID 5.01 dataset, mentioned above (Maoz et al., 2019). The most notable difference is that the data are presented in dyad years, instead of on the system level. This allows for easier transformation into the dyad month format required in this research, and more data points to use. That argument, combined with the advantages of the underlying CoW MID dataset make this the dataset of choice. For comparison among the CoW datasets, the table below presents the distribution of escalation classes among the three datasets discussed.

Table A-1. Usable cases CoW datasets

Level of escalation	No. of cases		
	MIDA (disputes)	MIDI (incidents)	MID Dyadic (dyad years)
Threat to use force	9	157	88
Display of force	228	2052	888
Use of force	350	1816	1234
War	4	8	92
Total	591	4033	2302

International Crisis Behaviour, version 14.0 (ICB)

The ICB codes 257 crisis episodes from 1918 to 2017 (Brecher et al., 2021; Brecher & Wilkenfeld, 2000). However, only 74 of those episodes took place from 1995 onwards. Furthermore, it includes a variable coding the level of violence, which could be used as an escalation measure. The dataset is though, very thinly populated.

Table A-2. Usable cases ICB dataset

Level of escalation	*No. of cases*
No violence	20
Minor clash	29
Serious clashes	14
Full-scale war	11
Total	74

ICB events, version 1.0 (ICBe)

The ICBe contains the same episodes as the ICB dataset, but contains a much more detailed description of the intra-crisis dynamics. It offers the best possibilities for measuring escalation, for example using the variables *fatalities* and *location* (Douglass et al., 2022). However, although it has plenty of accurately coded events, for our purposes, the amount of crisis episodes, equal to the ICB dataset, is quite low. Additionally, it offers far more granular access to conflict events than is required for the target variable. While a great asset when analysing specific conflicts, using this measure will first require aggregation of the specific events into dyad month. Although manageable, this would create an extra strain on the project resources.

UCDP Georeferenced Event Dataset (GED Global), version 21.1

The UCDP GED dataset is another example of a well-filled accurate event dataset, which has gained prominence in recent years (Pettersson et al., 2021; Sundberg & Melander, 2013). It contains a total of 261864 events, of which 99% have an associated starting date. However, interstate conflict falls under the wider category of *state-based* events, a coding decision which makes it impossible to filter out interstate conflict.

UCDP/PRIO Armed Conflict, version 21.1

The UCDP/PRIO AC is another contender. It contains measures for escalation, and codes explicitly for interstate conflict (N. P. Gleditsch et al., 2002; Pettersson et al., 2021). Unfortunately, it contains only fifteen useful cases, representing a mere ten unique conflicts with a specific starting month from 1995 onwards.

Issue Correlates of War, version 1.01 (ICOW)

The ICOW project (Frederick et al., 2017) offers data on salient issues in interstate conflict, of which only the territorial issues dataset covers the entire world. The datasets are based on the dyadic structure of the CoW MID datasets. It measures whether violence was used in resolving a territorial claim, but unfortunately, since it only focuses on territorial claims, we cannot use it for the prediction of the full spectre of interstate conflict, which may take place outside territorial claims. Additionally, it only contains data up to 2001.

Conflict Barometer - Heidelberg Institute International Conflict Relations Research 2020 (HIIK)

Finally, the HIIK Conflict Barometer datasets are another recently published dataset (Heidelberg Institute International Conflict Relations Research, 2021). Unfortunately, although it keeps track of conflict intensity, it contains only twenty-four instances of interstate conflict after 1995, which is far below the required hundreds of cases.

Table A-3. Summary of candidate datasets

	Codes for interstate conflict	*Relevant cases*	*Time period*	*Monthly (dis)-aggregation possible*	*Measure of escalation possible*
MID Data 5.0	Yes	591 MIDs, system level	1816-2014	Yes	Yes
MID Incidents 5.0	Yes	4033 incidents, system level	1992-2014	Yes	Yes
MID Dyadic Dataset 4.02	Yes	1651 undirected dyad years	1816-2014	Yes	Yes
MID Interstate War Data 4.0	Yes	24 wars, directed dyads	1816-2014	Yes	No
ICB V14	Yes	74 crises, system level	1918-2017	Yes	Yes
ICBe 1.1	Yes	58761 events	1918-2017	Yes, but difficult.	Yes, but difficult.
UCDP GED 21.1	No	176861 events, system level	1989 - 2020	Yes	No
UCDP/PRIO Armed Conflict 21.1	Yes	15 undirected dyad years	1946-2020	Yes	Yes
ICOW Territorial Claims 1.01	No	830 dyad years	1816 - 2001	Yes	Yes
HIIK 2021	Yes	24 wars, system level	1825 - 2020	No	Yes

Thus, the dataset meeting all four conditions best is the MID Dyadic dataset, with the number of usable cases setting it apart from most other datasets, and its dyad year format making it the better choice vis-à-vis the MIDA dataset.

A.1.2 Included countries

The countries included in the research are referred to as relevant states. There are 187 relevant states, which existed for the full period of time between 1995 and 2014, according to the CoW State System Membership dataset. However, there are eight states that came into existence after 1995. These were discarded to simplify the data manipulation process. If the number of dyads were allowed to fluctuate through time, the creation of a dyadic dataset, and adding all subsequent input variables to it would have been greatly complicated. For more details on the data manipulation process, see the file Data-manipulation_structural.ipynb in the supporting materials folder.

Of these eight excluded countries, four have been involved in an MID, for a total of fourteen MIDs. These are East Timor, Kiribati, Kosovo (2 MIDs) Montenegro (8 MIDs), Nauru, Palau (1 MID), South Sudan (3 MIDs), Tonga, and Tuvalu.

Furthermore, there are 26 duplicates in the CoW state list. These represent national 'rebirths', such as a liberation from occupation (Correlates of War Project, 2017). In some cases, this can be a long period of time (e.g., Latvia did not exist as an independent nation between 1940 and 1991). Fortunately, all of these instances were before 1995, which means they do not affect this project's list of relevant states.

Then, one country identified in the MID Dyadic dataset was not found in the state membership data. Uganda is noted as UGD, which should be the correct CoW abbreviation for Uganda, UGA. The four occurrences of this code were converted to the standard.

Finally, this leaves us with the 187 relevant countries:

Afghanistan
Albania
Algeria
Andorra
Angola
Antigua & Barbuda
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bhutan
Bolivia
Bosnia and Herzegovina
Botswana
Brazil
Brunei
Bulgaria
Burkina Faso
Burundi
Cambodia
Cameroon
Canada
Cape Verde
Central African Republic
Chad

Chile
China
Colombia
Comoros
Congo
Costa Rica
Croatia
Cuba
Cyprus
Czech Republic
Democratic Republic of the Congo
Denmark
Djibouti
Dominica
Dominican Republic
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Federated States of Micronesia
Fiji
Finland
France
Gabon
Gambia
Georgia
Germany
Ghana
Greece
Grenada
Guatemala
Guinea
Guinea-Bissau
Guyana
Haiti
Honduras
Hungary
Iceland
India
Indonesia
Iran
Iraq
Ireland
Israel
Italy
Ivory Coast
Jamaica
Japan
Jordan
Kazakhstan
Kenya
Kuwait
Kyrgyzstan
Laos
Latvia
Lebanon
Lesotho
Liberia
Libya
Liechtenstein
Lithuania
Luxembourg
Macedonia
Madagascar
Malawi
Malaysia
Maldives
Mali
Malta
Marshall Islands
Mauritania
Mauritius
Mexico
Moldova
Monaco
Mongolia
Morocco
Mozambique
Myanmar
Namibia
Nepal
Netherlands
New Zealand
Nicaragua
Niger
Nigeria
North Korea
Norway
Oman
Pakistan
Palau
Panama
Papua New Guinea
Paraguay
Peru
Philippines
Poland
Portugal
Qatar
Romania
Russia
Rwanda
Samoa
San Marino
Sao Tome and Principe
Saudi Arabia
Senegal
Seychelles
Sierra Leone
Singapore
Slovakia
Slovenia
Solomon Islands
Somalia
South Africa
South Korea
Spain
Sri Lanka
St. Kitts and Nevis
St. Lucia
St. Vincent and the Grenadines
Sudan
Suriname
Swaziland
Sweden
Switzerland
Syria
Taiwan
Tajikistan
Tanzania
Thailand
Togo
Trinidad and Tobago
Tunisia
Turkey
Turkmenistan
Uganda
Ukraine
United Arab Emirates
United Kingdom
United States of America
Uruguay
Uzbekistan
Vanuatu
Venezuela
Vietnam
Yemen
Yugoslavia*
Zambia
Zimbabwe

* Yugoslavia refers to the pre-2006 Serbia and Montenegro federation, and the post-2006 state Serbia. The state Montenegro was excluded. The name Yugoslavia was kept for the entire period 1995-2014, following Correlates of War.

A.2 Input data – Structural variables

This section presents a detailed overview of the input variables used in the models. All are coded as binary variables, except for inter-capital distances and dyadic peace years, which are interval variables. All possible dyad months were assigned a value for these variables. These variables were selected based on a review of the literature on the causes of interstate conflict.

For every variable, first, the concept it codes for is explained. Then, the grounds on which the variable was included are discussed. Subsequently, each subsection presents the operationalisation and data source of the variable, as well as finally, any remarks on data structure or data manipulation. An overview of the variables is given in the table below.

Table A-4. Structural variables

	Variable	*Operationalisation*	*Data source*
Political			
	Existing rivalries	1 if two states see each other as a significant threat, else 0	Thompson rivalry dataset, v2022
	Joint democracy	1 if two states score 6 or higher on the 0-10 Polity democracy scale, else 0	Polity5 Project
	Major power	1 if one state is classified a major power, else 0	CoW State System Membership dataset, v2016
Security			
	Alliance (broad)	1 if a any military alliance exists on the dyad, else 0	CoW Formal Alliances v4.1
	Alliance (defence pact)	1 if a defence pact exists on the dyad, else 0	CoW Formal Alliances v4.1
	Nuclear weapons (both)	1 if both dyad states possess nuclear weapons, and 0 otherwise	Nuclear Weapons – Our World in Data
	Nuclear weapons (one)	1 if only one dyad states possesses nuclear weapons, and 0 otherwise	Nuclear Weapons – Our World in Data
	Dyadic peace years	The number of years since the dyad was last involved in an MID	CoW MID Dataset
Geographical			
	Contiguity	1 if contiguous on land, or separated by less than 24 miles of water, 0 otherwise	CoW Direct Contiguity v3.2
	Distance	Distance between the dyad capitals in kilometres	EUGene
	Regions	1 when both states are in the same region, and 0 otherwise	World Bank

A.2.1.1 Existing rivalries

Existing rivalries refer to state pairs that see each other as a significant threat, sometimes for lengthy periods of time. These perceptions lead to tense relations, a heightened sense of insecurity and a higher risk of conflict, which is why rivalries are used in conflict prediction research (Braithwaite & Lemke, 2011).

Since the emergence of research on interstate rivalries in the 1990s, two data collection efforts have informed the mainstay of rivalry research. The first of these is the Diehl & Goertz (2000) dataset, which has received several updates and which now goes by the name of the Peace data (Diehl et al., 2021). This dataset contains information on an ordinal scale of interstate relations, ranging from 'severe rivalry' to 'positive peace,' and is based on six measurable underlying concepts, of which one is the frequency of MIDs on that specific dyad. The other dataset was developed by Thompson (2001), and also recently updated (Thompson et al., 2022). This dataset divides rivalries into positional, spatial, ideological, and interventionary rivalries, and is not based on measurable inputs, but on a careful examination of the historical record. Within rivalry research, these two datasets represent different camps in how to approach rivalry coding. Both are highly acclaimed and could be used for this project without problems. The Thompson dataset has small advantage, however. The Diehl et al. dataset is partly based on a frequency count of MIDs, which is quite similar to the dyadic peace year count used as a separate variable in this project, and the MID

frequency count itself may be easily added to this project as a variable in future research. So, based on the ease of future additions and variable diversity, this project will use the 2022 update of the Thompson dataset (Thompson et al., 2022), which qualifies interstate rivalries as two states that "regard each other as competitive or operating more or less in the same league" (Thompson et al., 2022, p. 2). These dyad instances are coded as 1, the rest as 0.

A.2.1.2 Joint democracy

The idea that democratic states are reluctant to go to war with each other is an old idea in the international relations field, and a central tenet of the democratic peace theory. It is supported by a wide of range of empirical studies (see, for example, Bremer, 1992; Maoz & Russett, 1993; Russett et al., 1995), which forms a robust ensemble underpinning the correlation between democratic dyads and peace (Kinsella, 2005; Hegre, 2014) (Kinsella 2005, Hegre, 2014). Nevertheless, plenty of debate persists regarding the causal pathways underlying this correlation (Hegre, 2014), with the theory unable to provide definitive answers. To illustrate, some competing explanations are set out. One explanation, tracing back to Kantian normative philosophy, comprises the assumption that states externalize norms of political behaviour, and thus are reluctant to solve issues by force (Maoz & Russett, 1993). Another explanation is that democratic leaders are more constrained in their use of force to the political structure of their states (Maoz & Russett, 1993; Hegre et al., 2020). Yet another posits that it are various socio-economic conditions that explain both democracy, and peace (Hegre, 2014). Still, none of these explanations enjoys scientific consensus. Moreover, some raise doubt that the countries making up the democratic peace will, in the future, continue to showcase the characteristics that cause this peace (Hobson, 2017), whatever these characteristics may be, thus reducing the future strength of the correlation. Nevertheless, considering the debate described above, the democratic peace remains a strong empirical finding and should find its way as a predictor in this model.

Based on this empirical finding, it is expected that a jointly democratic dyad is unlikely to experience MIDs. Yet, the effect of joint democracy on the level of escalation is less clear. One might expect an adverse effect, since democratic states would resolve their disputes peacefully, or on the other hand, one might expect that once democratic states do come into conflict, their conflict must be so severe that even their regime type could not prevent it, and thus a contributive effect on escalation. Braithwaite et al. (2011) found that the latter is more likely to be true, but their statistical findings merely point in the direction of this idea, they are not enough to prove it. Therefore, the effect on escalation is not clear on beforehand.

Operationalisation

The variable joint democracy uses data from the Polity5 Project by the Center of Systemic Peace (Marshall et al., 2020). This living data effort is the golden standard for measuring regime characteristics in quantitative international relations research and offers a composite variable coding for the strength of a state's democracy. A dyad is coded 1 if both states score 6 or higher on this 0-10 democracy scale, and 0 otherwise.

An advantage of this approach is that we can avoid using the problematic middle values of a combined autocracy-democracy score that ranges between -10 and 10. As the authors of the dataset note in their codebook, presenting all regime types on a linear scale from perfect autocracy to perfect democracy risks simplifying their characteristics. Democratic and authoritarian aspects can co-exist, resulting in a wide array of mixed regime types, which are difficult to capture on a linear scale. Especially the middle-range values of this scale, often termed *anocracies,* might not be representable of a state's exact regime type. Since we are interested in democracies only, the specific democracy variable can be used. This project follows Reed (2000) by creating a binary variable with the threshold at 6.

A.2.1.3 Major power

The presence of major powers is another well-established predictor for interstate conflict onset (King & Zeng, 2001). Due to their international influence and presence, major powers have extensive interests abroad, and are thus more likely to come into conflict with another state (Reiter & Tillman, 2002). Therefore, it is expected to correlate with conflict onset. There is no prior expectation in this research of the correlation between major powers and escalation.

Like all other variables, this one is coded dyadically. If one of two dyad members classifies as a major power, the variable is 1, if not, it is 0. The data source for the list of major powers is the standard in the field, the CoW State System Membership dataset, v2016 (Correlates of War Project., 2017). The major powers in this project are the United States, United Kingdom, France, Germany, Russia, China, and Japan.

A.2.1.4 Alliance

Alliance is yet another much used variable in interstate conflict prediction (see, amongst others: Reed, 2000; Gartzke, 2007; Stodola et al., 2021). Alliances are expected to have a negative influence on both dispute onset and escalation, since both states are on friendly terms with each other.

The source data are the CoW Formal Alliance dataset, version 4.1 (Gibler, 2009), which recognises four types of alliances. Defence alliances refer to defence pacts, neutrality refers to neutrality pacts, nonaggression comprises a pact where states vow not to attack each other, and last, ententes refer to states that will consult one another in case of a crisis.

These different types are accounted for in two variables, both coded 1 if an applicable alliance exists, or 0 otherwise. The broad alliance variable encompasses all types of alliances, and the defence alliance variable refers to defence pacts only. The reason behind the creation of a separate defence pacts only variable is that defence pacts, the strongest form of alliances, might be correlated more strongly than the general alliances with conflict onset and escalation. Including defence pacts separately in the feature space might add additional possibility to the modeling algorithms to extract relations between features and the target variables.

Data manipulation

Due to the format of the alliance data, special attention has been paid to duplicate values. The original data were collected at the alliance level, while the desired format is the dyad month level. This means that multiple alliances can exist next to each other in the same dyad month. An example is the alliance history of the United Kingdom and Portugal, who share three separate alliances.

Table A-5 Data excerpt showing UK-Portugal alliances

index	state1	state2	start year	end year	defence	Neutra-lity	nonaggression	entente	year
196	United Kingdom	Portugal	1816	2012	1	0	1	0	2012
43086	United Kingdom	Portugal	1988	2012	1	0	0	1	2012
47162	United Kingdom	Portugal	1949	2012	1	0	1	0	2012

When coding such cases at the dyad month level, which is required in this project, they create multiple entries per dyad month. To manage these duplicates, two operations are performed.

1. Drop duplicate rows where the treaty type is similar. Either both defence pacts, or both non-defence pacts.
2. In case of two dissimilar treaties, drop the non-defence pacts row. A defence pact is the most encompassing deepest of alliance, and presupposes the other types.

In the case of the UK and Portugal, this means that all dyad months will be coded 1 for both Alliance defence, and Alliance broad.

Finally, since the alliance data stretch back until 2012, dyad months after 2012 were assigned NaN values. Out of 9736 values used for the modeling dataset, 1009, or 10.4% were missing values.

A.2.1.5 Nuclear weapons

Nuclear deterrence theory posits that nuclear weapons have a pacifying effect, due to the apocalyptic consequences of a nuclear war, a central notion to the doctrine of mutual assured destruction. Although this is a popular notion, and might very well be true, especially when both states possess nuclear weapons, there is no academic consensus on the effect of nuclear weapons on conflict onset and escalation. Although the direction and size of these effects, remain open to debate, they do warrant nuclear weapons possession's inclusion as a predictor in this study. Two variables were created, one for mutual nuclear weapons possession, and one for one-sided possession.

When both states on a dyad have nuclear weapons, disputes between them are expected not to escalate, due to the risk of the apocalyptic consequences of nuclear war. However, dispute onsets are expected to be more frequent, mainly because nuclear power is correlated with other predictive factors, such as being a major power, and contiguity (Bennett & Stam, 2004, p. 137). This variable is coded 1 if both dyad states possess nuclear weapons, and 0 otherwise.

When only one state has nuclear weapons, disputes are also expected to be more frequent, because the nuclear-armed state enjoys an advantage in knowing that its opponent could not retaliate at the highest level of violence - it enjoys escalation dominance, the ability to move a conflict up the escalation ladder at a time and setting of one's own choosing. Note, however, that it might be the case that states have developed nuclear weapons, precisely because they were embroiled in conflict, which could make both variables be positively correlated with conflict onset (Bennett & Stam, 2004, p. 136). This second variable is coded 1 if only one of the dyad states possesses nuclear weapons, and 0 if otherwise.

Finally, the data were provided by the Our World in Data website (Roser et al., 2013). Of the countries under consideration, only Gambia was not included. It has, though, never possessed nuclear weapons, and was treated as such.

A.2.1.6 Dyadic peace years

A history of conflict has consistently been found to be predictor of future conflict (Reed, 2000; King & Zeng, 2001; Braithwaite & Lemke, 2011). Country pairs that experienced conflict recently are more likely to experience new conflict than those that did not, a phenomenon termed temporal or duration dependence in the literature. This variable is operationalised as the number of years that no MID has occurred between two states, starting from 1816, or the creation of the youngest state. This is calculated from the same MID data that make up the MID cases. Dyads experiencing repeated conflict show low values for this variable and can be described as dangerous dyads.

Notes on data structure

Only the relevant countries are included, and relatively few dyads extend all the way back to 1816, as can be seen in Figure A-1. To illustrate the variable construction, Table A-6 shows a sample of the peace years count between the United States (2) and Vietnam (816) for the dyad years 1995 to 1997.

Table A-6. Sample dyadic peace years variable

index	*combi*	*year*	*dyadic_peaceyears*
12137	2,816	1995	42
12138	2,816	1996	43
12139	2,816	1997	44

The variable is expected to be a strong predictor especially at the ends of the distribution of its values. Those very few country pairs that did not experience conflict for over a hundred or so years are highly likely to remain peaceful, whereas those dyads that have zero or one dyadic peace years are expected to experience conflict again.

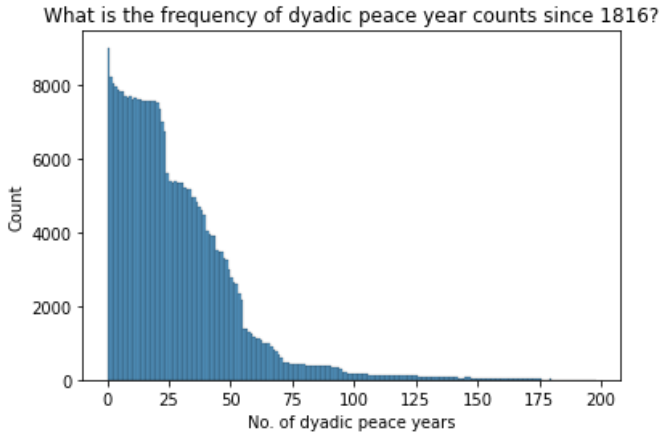


Figure A-1 Value counts of dyadic peace years variable

A.2.1.7 Contiguity

Contiguous states are more likely to come into conflict, as contiguity allows for intensive contact, both peaceful and conflictuous, between the two countries. This relation has been known for some time, see for example Maoz and Russet (1993) and Stodola et al. (2021). The source data for this variable are the CoW Direct Contiguity dataset, version 3.2 (Stinnett et al., 2002; Correlates of War Project, n.d.). Contiguity is operationalised dichotomously, with 1 if both states are contiguous on land, or separated by less than 24 miles of water, and with 0 otherwise.

A.2.1.8 Distance

Similar to contiguity, distance discourages conflict (Gartzke, 2007), because it limits the interactions of states, and makes extensive military conflict simple less feasible. Distance is operationalised as the distance between the dyad capitals in kilometres, using the EUGene software for data and data formatting (Bennett & Stam, 2000).

A.2.1.9 Regions

This is another measure of geographic proximity. A dyad is coded 1 when both states are in the same region, and 0 otherwise. The seven regions are: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, North America, South Asia, and Sub-Saharan Africa. The data are provided by the World Bank analytical grouping of countries (World Bank Group, 2018, p. x).

A.3 Multicollinearity

The following table presents all thirty-seven features out of 279 that have a variance inflation factor higher than 5, which indicates the presence of multicollinearity. All are event count variables.

Table A-7. VIF scores higher than 5

	Feature	*Variance inflation factor*
1	Appeal for easing of political dissent	∞
	Accede to requests or demands for political reform	∞
	Host a visit	20620.82
	Make a visit	20586.02
	Provide military protection or peacekeeping	390.01
	Receive deployment of peacekeepers	389.97
	Ease military blockade	341.84
	Engage in political dissent	316.07
	Use conventional military force	32.27
10	Make statement	19.51
	Use unconventional violence	16.97
	Employ aerial weapons	14.54
	Expel or withdraw peacekeepers	11.33
	Const - (constant used in VIF calculation)	11.20
	Occupy territory	10.56
	Appeal for easing of administrative sanctions	9.87
	Impose blockade, restrict movement	9.64
	fight with artillery and tanks	9.26
	Demobilize armed forces	9.25
20	Reject	8.83
	Impose administrative sanctions	8.37
	Abduct, hijack, or take hostage	8.29
	Give ultimatum	7.86
	Grant diplomatic recognition	7.51
	Make an appeal or request	7.33
	Express intent to de-escalate military engagement	7.24
	Threaten	7.00
	Accuse	6.85
	Demand	6.34
30	Refuse to ease administrative sanctions	6.29
	Retreat or surrender militarily	5.93
	Express intent to meet or negotiate	5.91
	Consult	5.91
	Mobilize or increase armed forces	5.51
	Criticize or denounce	5.43
37	Violate ceasefire	5.41

TU Delft

The Hague Centre for Strategic Studies