Quality & Quantity
https://doi.org/10.1007/s11135-019-00882-w

Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field

Felix Ettensperger[1]

Abstract
Machine learning algorithms and artificial neural networks promise a new and powerful approach for making better and more transferable predictions in global conflict research. In this paper, a novel conflict dataset for the prediction of conflict intensity is introduced. It includes seven socio-economic and political indicators spanning a set of 851 country-years. This set of indicators is combined with conflict intensity data covering the timeframe of 2009–2015 to build a viable predictor framework. With this dataset as a foundation, a wide range of different predictive methods are tested, including linear discriminant analysis, classification and regression trees, k-nearest neighbor, random forest and several series of advanced artificial neural networks including a novel non-sequential long-short-term memory setup. Acknowledging the potential of deep learning techniques for many disciplines and projects, this paper shows, that for this type of assembled medium sized data, resembling many common research frameworks in Social and Political Sciences, using neural networks as singular approach might not be fruitful. The advantages of neural networks do not always outweigh their practical and technical disadvantages in small or medium data settings. The argument derived from this study is that researchers should combine Supervised Learning Algorithms and Deep Learning Networks as a general approach in similar predictive setups, or carefully evaluate for each dataset and project if the added complexity accompanied with using networks is indeed translating into better predictive performance.

Keywords Forecasting · Machine learning · Random forest · Prediction · Conflict · Neural networks

1 Introduction

Beginning with the landmark studies of Collier and Hoeffler (1998, 2004), a quantitative turn in conflict research lead to a surge of publications studying quantitative risk factors affecting the conflict propensity of regions and states (Fearon and Laitin 2003; Hegre

✉ Felix Ettensperger
felix.ettensperger@politik.uni-freiburg.de

[1] Department of Political Science, Chair of Comparative Politics, Albert-Ludwigs-University Freiburg, Freiburg, Germany

Published online: 15 May 2019

and Sambanis 2006). From better understanding and measuring the effect of this variables, many political conflict researchers turned to the question how this knowledge can be applied to forecast global conflicts and how dangerous conflict developments can be identified beforehand, partially contained or even entirely prevented.

Research teams world-wide are putting great effort into developing different prototypes of predictor frameworks that can correctly and timely forecast conflict escalation in the future based on the economic, political, demographic and social data available today in individual states and regions (Schmeidl and Bond 2000; Hudson et al. 2008; O'Brien 2010; Goldstone et al. 2010; Smidt et al. 2016).

Despite the efforts to translate accessible risk variables into a viable framework to predict conflicts and effectively forecast future risks, many studies and projects encounter it challenging to build a framework sensible and flexible enough to correctly predict the immense complexity of global conflict development and forecast with high accuracy and precision (Cederman and Weidmann 2017: 474).

While collecting and applying more and better data can be a solution for the future, this contribution proposes and compares new machine-learning algorithms and neural networks as methodological addition to the toolkit of conflict research to improve predictive quality of existing frameworks. At the moment many research projects are still based exclusively on conventional regression methods that are often inflexible in predicting the complex nonlinear interactions of different variables and contexts and thus potentially less accurate in predicting future conflict escalation than modern learning algorithms and networks.

Several Machine Learning techniques have been applied and tested for conflict prediction in isolation before (Beck et al. 2000; Perry 2013; Colaresi and Mahmood 2017) but a comprehensive comparison and evaluation of their potential in the field is still missing.

The objective of this contribution is to test and compare systematically the accuracy of various different machine-learning algorithms and networks, including novel and modern techniques used for the first time in conflict prediction, for example Artificial Neural Networks with complex LSTM Layers.

Introducing and sampling seven socio-economic indicators, this paper subsequently applies a total of eleven different techniques of prediction, including k-nearest neighbor (k-NN), random forest (RF), feed-forward neural networks (FFNN) and recursive neural networks with long-short-term-memory (LSTM) layers. The accuracy of these techniques in conflict prediction is compared to two linear regression-based categorization methods as baseline models. The contained small-scale conflict prediction framework is suitable for predictions with supervised learning algorithms and deep learning and at the same time resembles many existing long-term conflict prediction setups applied globally.

Technical comparisons of Machine learning techniques and networks have been conducted and published before (Caruana and Niculescu-Mizil 2006), but this paper focuses specifically on socio-economic and political data and highlights the advantage of applying different machine learning techniques simultaneously in conflict prediction frameworks. The publication concludes with recommendations regarding the sensible application of artificial intelligence in the realm of conflict prediction and how to establish a comprehensive interdisciplinary research methodology in regard to conflict forecasting.

2 Literature review of political conflict prediction

Political conflict prediction is a highly diverse field. Expert forecasting, simulation models and econometric approaches are applied simultaneously in various different projects (Chadefaux 2017). Some approaches focus on political event data as a basis for prediction by use media content and verbal statements of political actors to forecast the impending escalation of conflicts in regions and states (Schmeidl and Bond 2000; Hudson et al. 2008). Other approaches concentrate on econometric data and the political background of states and regions to predict the conflict propensity and future risk levels of states. Both research paradigms carry advantages and disadvantages. Some new approaches combine both sides to improve prediction further (O'Brien 2010).

In this paper a prediction framework is presented for predicting state failure and state-wide conflict escalation based on socio-economic, demographic and political data. The proposed framework is highly comparable to the currently applied approaches of the Political Instability Task Force (PITF) or the Global Conflict Risk Index (GCRI) of the European Union (Goldstone et al. 2010; Smidt et al. 2016).

It also resembles the underlying macroeconomic framework of the Integrated Crisis Early Warning System (ICEWS) currently developed for the U.S. Government (O'Brien 2010). All of these mentioned frameworks use aggregated data on the state-level to predict conflict levels in the near- and medium future. Most of the introduced frameworks however still exclusively apply linear, variance based prediction methods, for example logistic regression (King and Zeng 2001) for predicting conflict deterioration.

In recent publications, advanced Machine Learning (ML) techniques have been applied to conflict prediction and research (Beck et al. 2000; Perry 2013; Muchlinski et al. 2016) and already strategies and frameworks for improving ML settings in Conflict Research are emerging and are tested with new algorithm-based approaches (Colaresi and Mahmood 2017).

Muchlinski and his colleagues applied Random Forest algorithms to generate predictions of civil war outbreak based on three landmark studies and the therein contained variables (Muchlinski et al. 2016: 93; Fearon and Laitin 2003; Collier and Hoeffler 2004; Hegre and Sambanis 2006). This series of tests revealed, that despite high p-values in the corresponding regression analysis and the well-tested and controlled application of these techniques in the former studies, the significant variables extracted from the studies showed the tendency to not always consistently and reliable improve prediction quality in a machine learning design (Muchlinski et al. 2016: 98). This aligns with previous results of Ward, Greenhill and Bakke, who observed the paradox behavior that adding more, significant variables sometimes even reduced the prediction quality if added in a forecasting setting (Ward et al. 2010: 367).

All these recent discoveries are generating an important discussion of predictive power versus significance and invite further research into new modes of prediction.

3 Advantages of using modern machine learning methods to forecast conflict

The advantages of using supervised learning algorithms and neural networks over regression models to identify patterns in data and predict categories are manifold and well established (Bishop 2006; Alpaydin 2004; Hastie et al. 2009): Many techniques can operate with more variables than cases, unlike most regression-based models (Segal 2003; Breiman 2001). There is no methodological limit to the size of data incorporated and integrating more data usually transfers into better training of predictors and eventually higher predictive quality. Thus, they are naturally excelling in big data environments, often incorporating millions of cases with hundredths or thousands of variables simultaneously (Goodfellow et al. 2016). They are flexible to easily incorporate non-linear data and can handle information that is strongly interdependent and multicollinear. Thus, they are in principle well suited for the highly interconnected socio-economic and political data available in the field of conflict prediction (Cederman and Weidmann 2017: 465; Hegre et al. 2016).

The success of ML techniques does however depend on the availability of sufficient data for training (Goodfellow et al. 2016: 19–20). As political, social, socio-economic and cultural data is relatively sparse and still more limited in scale compared to what is accessible in other disciplines, incorporating sufficient data might be one crucial limitation to the applicability of ML techniques in Political and Social Sciences and all related fields.

Supervised learning algorithms are in general considered less sensible to limited data than deep learning (network) setups. Thus, combining both applications might be the best way forward and more productive under these constraints. Additionally, networks confront us with so-called black box problems, as it remains not always clearly traceable, how the network reaches a prediction or how strong certain variables are weighted and considered (Hastie et al. 2009: 409; Benitez et al. 1997).

The uncertainty regarding the performance of socio-economic variables for conflict prediction warrants to include new variables and methods to be tested in innovative prediction frameworks. By testing the here presented set of socio-economic and political variables with a conflict intensity indicator that has not been tested before in ML settings, the Global Peace Index (GPI), this paper also adds diversity to the overall discussion of conflict prediction. But predominantly, this paper provides technical advancements by systematically comparing predictor frameworks and adding novel deep learning approaches like LSTM Networks to the methodological toolkit of conflict prediction.

The potential and quality of all prediction frameworks is intensively discussed by experts in the field and there remains a lot of skepticism regarding the potential of building and improving genuine long-term prediction frameworks with supervised learning techniques and networks (Cederman and Weidmann 2017). As this paper shows, a medium sized framework can indeed benefit from incorporating and using several ML techniques in parallel by testing performances of various approaches and comparing results. If this translates into better frameworks to build a working long-term predictor or a reliable conflict warning system of course remains to be seen.

4 Data and Variables

This paper compares different ML approaches to predict observations in a test sample based on a global conflict level dataset created with seven prominent and accessible socio-economic indicators. Structurally, the data is identically organized as many existing country-year conflict prediction frameworks, for example the Global Conflict Risk Index, a long-term prediction framework for global conflict escalation established by the Joint Research Centre of the European Union (Smidt et al. 2016).

The selection of the seven socio-economic indicators included in this compact framework is based on previous conflict research publications and the general conflict literature, established prediction frameworks and pre-test models as well as explorative studies regarding the correlation and significance of these measures in regression analysis.

The included indicators in the pattern learning framework are:

- The *Bayesian Corruption index* (Standaert 2015) reflecting on the effect of corruption on conflict development (see also Neudorfer and Theuerkauf 2014; Le Billon 2003).
- *GDP* as fundamental economic performance benchmark (related to conflict risk see also Fearon and Laitin 2003; Collier and Hoeffler 2004).
- The *freedom house index variable for electoral process* measuring the degree of free and fair elections provided by the political system (Vreeland 2008; Collier and Hoeffler 2005).
- *Uneven economic development indicator* from the Fund for Peace Institute measuring uneven economic development expressed through uneven distribution of incomes and services in a state (FFP 2017; for inequality as cause for conflict see Cramer 2003; Stewart 2008).
- *Indicator of economic globalization* established by Axel Dreher in the KOF Globalization Index (Dreher 2006; for conflict and trade see Dorussen 1999; Hegre 2002).
- *Freedom of the press score* provided by Freedom House is the sixth indicator measuring restrictions, pressures and controls applied to members and institutions of the press (see Gohdes and Carey 2017).
- *Refugee population by country of Asylum* provided by the Fund for Peace. This indicator measures the extend of pressure generated due to population inflow, refugee population, extend of refugee camps and strains on the absorption capacity for displaced persons within a country (see for conflict and refugee inflows: Feder 1998; Ek and Karadawi 1991).

To reach a viable amount of cases for ML techniques each case does not represent a singular country but a country-year unit. Over 6 years, in a timeframe between 2009 and 2014, these country-year are accumulated to form a dataset of 851 observations. The advantage of year-by-year accumulation is that enough data becomes available to reliably use ML techniques in this setup.

A potential disadvantage of this approach is that a high degree of autocorrelation between years is incorporated and affecting prediction accuracy positively. By separating the test and training set in various combinations this effect can be evaluated. Furthermore, by manipulating single variables in observations and predicting them, the influence of autocorrelation can be estimated. More details in regard to autocorrelation tests are presented in the "Appendix".

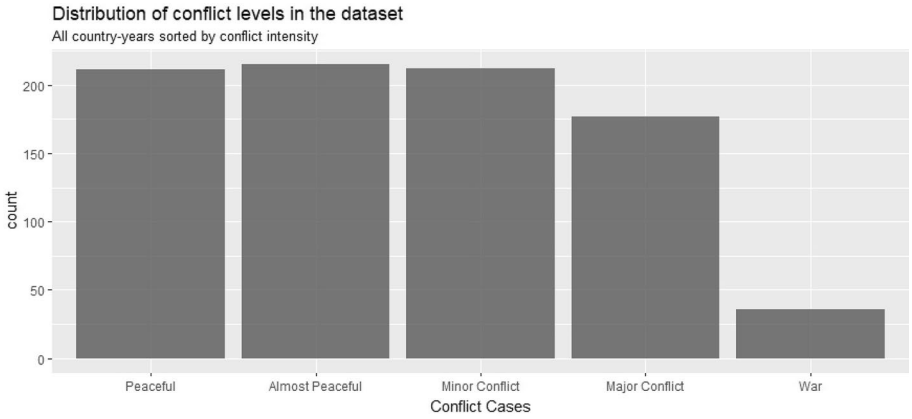

Fig. 1 Data set composition—distribution of 851 country-years by conflict level category

Some methods that are applied in the following prediction setup require a categorical variable as target. Thus, all GPI scores are transferred to a categorical system and systematically compared in the main part of the paper. For the application of machine-learning with linear quasi-metric target-variables, additional information is provided in the "Appendix".

To form conflict intensity categories, the GPI Score, a linear quasi-metric variable, is based on mathematical thresholds split into five different categories. There is a 1-year lag in the model between the political and socio-economic indicators and the conflict variable, to exclude and control for short-term interaction effects of conflict on the training variables. Each country-year is sorted into these five categories according to its GPI Score, reflecting the overall level of violence present in the corresponding state in the following calendar year.

The five GPI-Categories formed from the linear data are: Peaceful, Almost Peaceful, Minor Conflict, Major Conflict, War.

The variable distribution of the 851 country-years is made visible in the included distribution- and table plot provided in Figs. 1 and 2 which is sorted by conflict intensity. The visualization of the variable distribution in the table plot (Fig. 2) is already indicating a substantial relationship between the socio-economic indicators and the conflict variable.

In the "Appendix" a short discussion on the categorization is provided and an iteration of machine learning prediction with the quasi-metric original conflict data is used in a random forest regression trees model and compared to nested regression results.

There are two concerns with the data setup that might be considered and are therefore addressed: First the unorthodox application of GPI Data as conflict variable in this framework is a novel approach. The GPI is a composite index of 24 different indicators, not all of them directly related to the security situation prevalent within a state (for example weapons exports or state possession of nuclear weapons). To use the GPI as data foundation has however a lot of upsides, for example it provides a good global coverage of cases within a short timeframe of only several years, which provides high consistency in the data, which is one of the most important criteria for the predictor to effectively create meaningful predictions.

The GPI Score also does effectively differentiate between peaceful and slightly lesser peaceful countries and not only describes violent conflict incidents or violent deaths in war

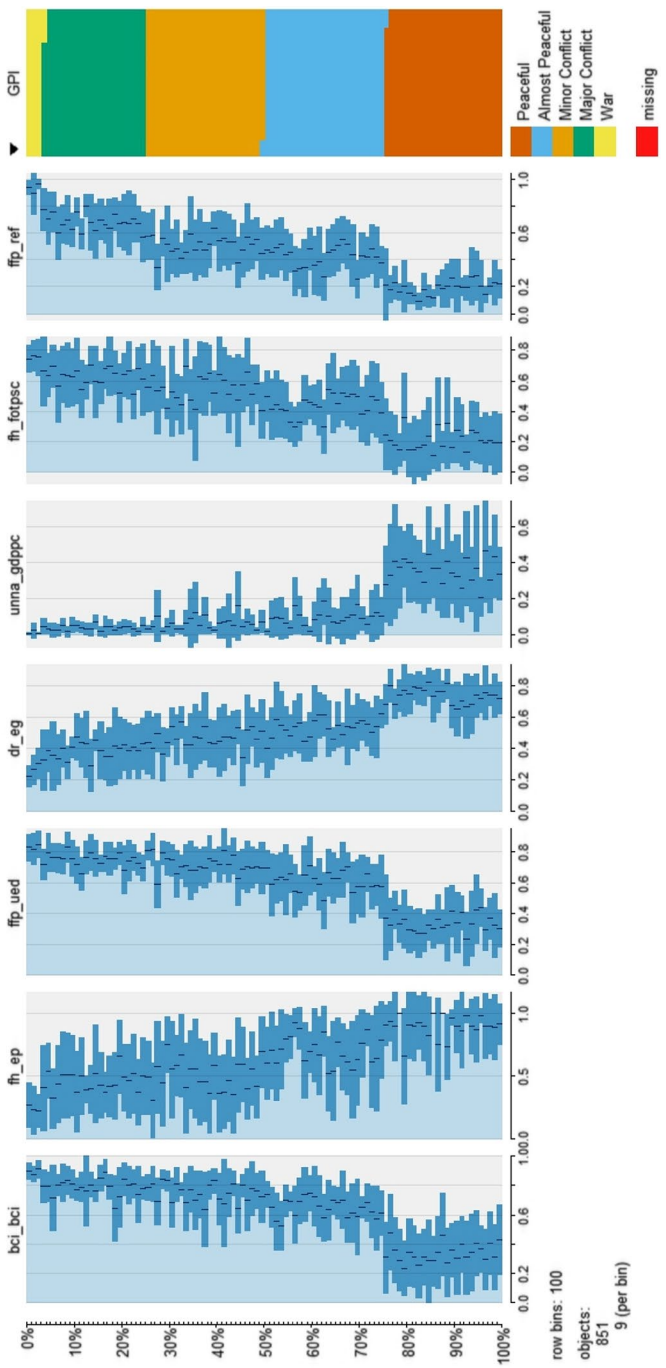

Fig. 2 Table plot—distribution of 7 socio-economic indicators sorted by conflict category

or war-like settings. It furthermore included crime indicators and thus incorporates the difficult relationship between excessive organized crime and peacefulness in countries.

Testing this framework with cleaned data, excluding all non-relevant indicators like gun ownership or nuclear armament, revealed that the effect of these included variables, however disturbing they might appear to the objective of conflict prediction, is comparatively small. Because of the 5-category classification framework which is put on top of the data, cleaning a test sample and comparing its results with the raw data leads in less than 0.5% of country-years to an effective change between the cleaned and uncleaned setup (and always only a change of maximum one category). A work- and time-intensive modification of all 851 country-years, also raising potential questions of the coding and inclusion of conflict items, was thus discarded.

A second objection might be that conflict is complex and reducing it to a categorical variable in a country-level framework is always excluding a lot of empirical information. Conflicts can be separated into different types of ethnical, political, religious or cultural disputes. They can be categorized as inner-state, intra-state or extra-state types of conflict or hybrid types of escalation events (Sarkees et al. 2003). The mixing of inter- and inner-state conflict events might be a problem if a deeper understanding of each phenomenon is the overall objective. Yet, from a prediction perspective, the goal of predicting conflict intensity in a country does not require to build a separate framework for each specific type of conflict.

The intention in this contribution is, that by allowing all conflict incidents and types together, the algorithm, not the researcher, has to adapt and learn how to identify specific symptoms and conflict clusters within the socio-economic and political indicator patterns pointing to inter-state or inner-state or any mixed or hybrid form of conflict. The predictor is furthermore not intended to be a singular, deterministic tool for prediction. The framework might be used in combination with other methods and studies to reveal long-term socio-economic and political conflict tendencies and simulate potential developments based on socio-economic data forecasts.

5 Predictor setup

To build a viable predictor for future conflict levels, testing beforehand how well the predictor performs on available data is essential.

In each of the following test series 75% of all country-years are used to train a predictor. After training the algorithm, network or regression model, the remaining 25% of the 851 observations (210 observations) are predicted using only the socio-economic data of each country-year.

The principal benchmark for the quality of the prediction is “accuracy” (in contrast to the root mean squared error in regression). The reported accuracy of each iteration in the next chapters constitutes the instances of correct classifications into the real category the case really belonged to. A value of 0.7 translates into 70% correct identification to the corresponding conflict level. Furthermore, a correlation matrix is provided to easily evaluate how far wrong predictions deviated from the real conflict category for different techniques.

Kappa, the second given value in LDA, CART, k-NN and RF models, is classification accuracy normalized by class imbalance. A useful measure to evaluate very asymmetric class distributions. It is another common benchmark in machine learning techniques and often reported alongside accuracy.

The loss value in network models is the mean squared error of the network function and is minimized during the epochs of the subsequent learning process.

Given that five categories are incorporated and applied in this setup; the minimum accuracy we would expect from simply guessing the corresponding category for each case in the test sample is overall 0.2 or 20% (Depending on the category, the chance of correctly assigning a case might vary due to different category sizes). If our variables and the applied techniques are working well for conflict prediction, we should observe considerably higher accuracy in the test sample.

Three different result types are generated and evaluated for each technique:

1. *K-fold cross-validation (5 CV)* 210 Country-years are randomly selected to the test sample. The resampling is repeated 5 times in so-called cross-validations to control for specific effects related to the random selection of country-years. The remaining 641 country-years are used each time to train the applied algorithm or network. After the training process, the 210 excluded observations are predicted by the trained framework and the accuracies of these predictions are compared to the real conflict intensity prevalent in each case.
2. *Predicting all country-years of 2013 and 2014 based on previous data* In the first year-wise iteration only the data from the years 2009–2012 (561 country-years) is used to train the framework and predict the test sample of only the year 2013 observations (145 country-years). In the second iteration the 2009–2013 data (706 country-years) is used to predict the conflict levels of all country-years of 2014 (145 country-years). Due to the increase in learning data for the 2014 prediction, the last year should be featuring consistently higher accuracy than 2013.
3. *Excluding entire countries from the training sample* The last applied test series is, to prevent the explicit learning of a selected list of entire countries. This is done by separating the data frame along country lines in several series of cross-validations. The performance in this controlled test series is showing the performance of the predictive technique with incomplete learning.

If an efficient ML algorithm or network learns the variable structure of e.g. Switzerland-2010, most probably the data from the very similar country-year pattern Switzerland-2011 can be easily predicted due to the almost identical variable structure between country-years in the normal training series. By controlling for this, separating countries in block, it is possible to estimate how solid the prediction technique is responding to completely new, unfamiliar data inputs. This is a very disruptive approach for machine learning techniques as it foils to a high degree their strength of pattern recognition. It provides us with a benchmark to show the dependency of data similarity for correct predictions. Additionally, it is interesting to detect overfitting in network iterations, as they tend to adapt too well to the training data, achieving high levels of accuracy on very similar data, but critically high error rates on more unfamiliar looking data. However, applying this strict separation, 25% of country data patterns never enters the autocorrelation training set in each iteration, we thus lose a lot of empirical diversity and information for the training process. Hence, the autocorrelation results should be considered with caution and always regarded as a low boundary of potential accuracy with completely unfamiliar patterns of data.

Table 1 Overview of all tested algorithms and methods

Name	Type
Linear discriminant model (LDA)	linear, variance-based algorithm
Classification and regression tree (CART)	linear, regression-based algorithm
Support vector machines (SVM)	non-probabilistic linear classifier algorithm
Bagged tree model	ensemble learning, tree-based algorithm
C 5.0 classification tree	ensemble learning, tree-based algorithm
k-nearest neighbour (k-NN)	non-parametric classification algorithm
Random forest (RF)	ensemble learning, tree-based algorithm
Feedforward neural network (1-layer)	artificial neural network (ANN)
Feedforward neural network (3-layer w/dropout)	artificial neural network (ANN)
LSTM neural network (200 LSTM nodes)	recurrent artificial neural network (RNN)
LSTM neural network (300 LSTM nodes)	recurrent artificial neural network (RNN)

Table 2 Linear Discriminant Model

5-cross validations *LDA—maximum likelihood*	Accuracy on the test sample	Kappa on the test sample
Iteration 1	0.5238	0.3733
Iteration 2	0.5381	0.3909
Iteration 3	0.5952	0.4677
Iteration 4	0.5143	0.3583
Iteration 5	0.5286	0.3807
Predict year data *LDA—max. likelihood*	Accuracy on the test sample	Kappa on the test sample
Iteration 2013	0.5034	0.3544
Iteration 2014	0.5034	0.3538
Exclusion test *LDA—max. likelihood*	Accuracy on the test sample	Kappa on the test sample
Iteration A	0.4044	0.2109
Iteration B	0.5070	0.3513

6 Results

LDA, CART, k-NN, RF, FFNNs and LSTM results are presented in detail in the following paragraphs. A complete comparison of the performance of all 11 algorithm and methods is provided in the summary section. All tested algorithms and models are listed in Table 1.

Table 3 Distribution of predictions in the LDA iteration—prediction of 2014 country-years

LDA 2014	Real				
Prediction	Peaceful	Almost peaceful	Minor conflict	Major conflict	War
Peaceful	**27**	8	1	0	0
Almost peaceful	10	**14**	9	5	0
Minor conflict	1	8	**13**	3	0
Major conflict	0	8	13	**17**	4
War	0	0	0	2	**2**

Table 4 CART model

5-cross validations *CART*	Accuracy on the test sample	Kappa on the test sample
Iteration 1	0.5810	0.4475
Iteration 2	0.4810	0.3222
Iteration 3	0.5143	0.3655
Iteration 4	0.5048	0.3440
Iteration 5	0.5286	0.3878
Predict year data *CART*	Accuracy on the test sample	Kappa on the test sample
Iteration 2013	0.5517	0.4065
Iteration 2014	0.4828	0.3124
Exclusion test *CART*	Accuracy on the test sample	Kappa on the test sample
Iteration A	0.4444	0.2598
Iteration B	0.5023	0.3408

7 Linear discriminant analysis (LDA) and classification and regression tree models (CART)

As a baseline for prediction accuracy a simple linear model is tested with this data to examine, how a variance-based model performs regarding to prediction quality with the proposed setup (Table 2).

With the 5-CV LDA results gathered in this test series a median accuracy of 54% can be established. For predicting the following year, the model performs stable at around 50%. This means 50% of observations in the test sample are correctly assigned to their true conflict category. 93% are assigned to the true and directly adjacent categories. The distribution of the prediction is shown in the cross-prediction matrix provided in Table 3.

Interestingly, this simple LDA techniques performing relatively stable in the exclusion test, as they do not improve their prediction from previously learned patterns in the training sample but weight variables linearly into the prediction. The LDA might still be a good

choice for prediction if completely unfamiliar, novel data inputs are considered and very high levels of abstraction are required.

As a second regression-based test, a Classification and Regression Tree (CART) model for class prediction is applied to similarly examine its performance (Table 4).

A standard CART model performs slightly less accurate that the linear LDA model. It shows additionally the unexpected behavior of achieving lower accuracy with more data between the year 2013 and 2014 predictions.

Both techniques manage to predict conflict categories right about 2.5 times better than a random assignment (Accuracy: 20% random prediction vs. 50% LDA/CART). Weighting the available information, they manage to directly assign observations correctly in about half of all instances. Only 7% of predictions are completely off and miss the right category by more than one category (vs. 40% expected in random setups).

Currently, both types of linear prediction models are still widely applied on a global scale in various prediction setups and thus their performance can be assumed to be a general baseline or reflection of the current status quo in prediction. Based on this accuracy we can evaluate how much better modern supervised algorithms and networks might perform in the following tests.

8 Supervised learning algorithms: k-nearest neighbor and random forest

Previous work from Muchlinski and colleagues (2016) has shown, that using supervised learning algorithms, especially Random Forest, for the prediction of conflict onset applying PRIO Data is working exceedingly well.

Random Forest and k-NN self-learning techniques are probabilistically approaches to pattern recognition and classification.

The k-NN algorithm determines new, unclassified cases based on the proximity of earlier learned neighboring cases. The assignment of categories with this technique is thus purely based on the similarity of new observations to previously learned observations. The k-NN algorithm offers, in terms of complexity in machine learning, a rather simple, instance-based approach for determining what category is most probable for a case based on the proximity to their relative environment of learned examples.

Random Forest, the second approach, is a technique based on the creation of multiple decision trees. The first forest related models were introduced by works of Amit and Geman (1997) as well as Ho (1995). The random forest technique was later established, refined and coined by Leo Breiman (Breiman 2001) and Cutler et al. (2012). A random forest model generates a "forest" of so called decision trees and applies during the creation of trees in this forest statistical "bagging" (*or* Bootstrap aggregating)—a procedure proposed by Breiman to reduce variance in an estimated prediction function (Breiman 1994, 1996). This procedure allows to generate a certain measure of controlled variance in this forest of de-correlated trees (Hastie et al. 2009: 587). In the end, all created decision trees conduct a majority vote determining the most probable categorical assignment to a specific case.

The presented algorithms surpass linear models by a significant degree, as can be demonstrated in Tables 5 and 6.

k-NN models regularly excel in environments were each class has many prototypes and the decision boundary is very irregular (Hastie et al. 2009: 465), a setup we would overall anticipate in context of the available conflict data. k-NN only evaluates the

Table 5 k-NN model

5-cross validations *k-NN algorithm*	Accuracy on the test sample	Kappa on the test sample
Iteration 1	0.7286	0.6435
Iteration 2	0.7238	0.6384
Iteration 3	0.7429	0.6624
Iteration 4	0.7048	0.6102
Iteration 5	0.7238	0.6376
Predict year data *k-NN algorithm*	Accuracy on the test sample	Kappa on the test sample
Iteration 2013	0.7655	0.6957
Iteration 2014	0.7379	0.6568
Exclusion test *k-NN algorithm*	Accuracy on the test sample	Kappa on the test sample
Iteration A	0.3422	0.1505
Iteration B	0.4225	0.2399

Table 6 RF model

5-cross validations *RF algorithm (500 trees)*	Accuracy on the test sample	Kappa on the test sample
Iteration 1	0.8048	0.7430
Iteration 2	0.7952	0.7314
Iteration 3	0.8048	0.7440
Iteration 4	0.8000	0.7367
Iteration 5	0.7762	0.7062
Predict year data *RF algorithm (500 trees)*	Accuracy on the test sample	Kappa on the test sample
Iteration 2013	0.7724	0.7039
Iteration 2014	0.8000	0.7380
Exclusion test *RF algorithm (500 trees)*	Accuracy on the test sample	Kappa on the test sample
Iteration A	0.4178	0.2335
Iteration B	0.5211	0.3698

distance and categorical similarity of neighboring cases to select a predicted category. The algorithm achieves accuracies of 73–76% in the year 2013 and 2014 iterations and only slightly less with 72–74% in the fivefold cross-validation. Yet, as it is dependent on the proximity of similar observations for prediction, it performs relatively bad on more abstract data as shown in the exclusion tests.

The iterations with Random Forest, a tree generating algorithm based on the works of Breiman (2001), tops the k-NN approach in prediction quality, in all series and modes (Table 6).

The algorithm assigns observations in the test sample with great success, reaching 80% of accuracy in the year 2014 series. It assigns the observations with high precision to the right category and if errors occur, they almost never exceed one category, as shown in the cross-prediction frame in Table 7.

The Random Forest Algorithm is a tree-based prediction technique. It generates a forest of randomized decision trees applying a complex measure (bootstrap aggregation) to randomize the tree generation process. After generating a forest of decision trees, they are pruned and then a majority vote on the most likely categorical outcome is conducted (Breiman 2001). This process seems to be especially suitable for the socio-economic and political data incorporated in the presented predictor framework. The overall direct accuracy of this approach is reaching 80% in the year 2014 series, about 4 times higher than by randomly selecting a category (20%). The wrongly predicted observations also cluster heavily around the right categories: if the predictor fails to assign an observation correctly, the assigning error only exceeds one category in 1% of all predictions. Furthermore, with 50% of accuracy in the exclusion test, it performs equally well to completely unfamiliar data as the LDA and much more consistently than a k-NN Algorithm.

Further advantages of the RF technique are, due to the mechanics involved, that overfitting is automatically corrected (Hastie et al. 2009: 587–588) and that tuning of hyperparameters is generally fast and simple. In many research frameworks extensive parameter tuning is not even necessary as a basic RF iteration with 500 trees already provides near-optimal results.

Examining the error rate in one random RF iteration, it seems clear that the number of trees in this framework rapidly leads to a stable equilibrium of prediction quality per category as seen in Fig. 3.

Beyond 200 trees the prediction quality for each category remains almost entirely stable, making further extensions to the forest size unnecessary.

Furthermore, complex calibration processes using Platt's method or Isotonic Regression often applied to improve SVM or Bagged tree results is rather inefficient and unnecessary with Random Forests (Caruana and Niculescu-Mizil 2006: 7). Another advantage of the RF Algorithm is its transparency regarding the individual importance of the included variables. The Variable Importance Measure (VIM) can reliably describe the influence of the Variables on the overall prediction accuracy, much like the *t* test in regression. In this example

Table 7 Distribution of Predictions in the RF Iteration 2014

RF 2014	Real				
Prediction	Peaceful	Almost peaceful	Minor conflict	Major conflict	War
Peaceful	**36**	2	0	0	0
Almost peaceful	2	**28**	3	2	0
Minor conflict	0	8	**27**	5	0
Major conflict	0	0	6	**20**	1
War	0	0	0	0	**5**

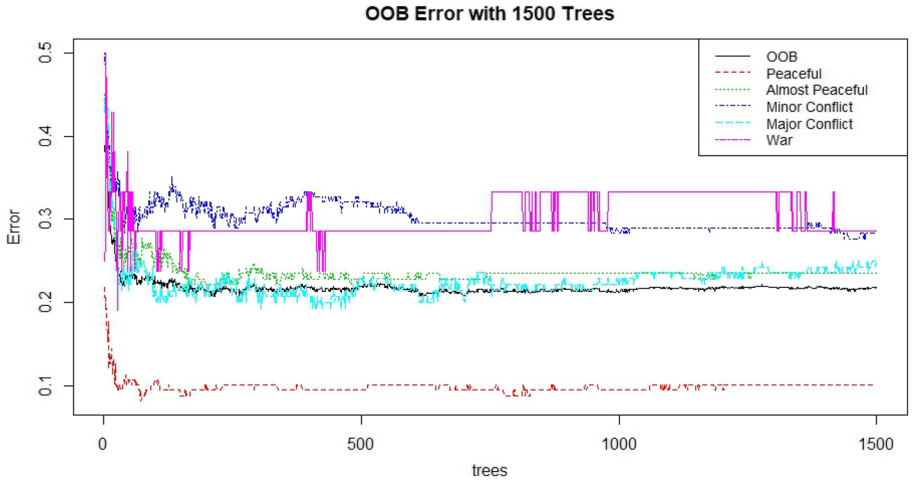

Fig. 3 OOB error of random forest iteration 1 with increase forest size

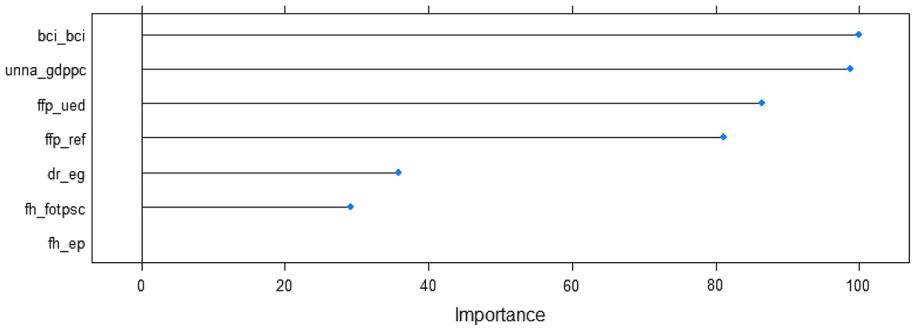

Fig. 4 Variable importance measure—500 tree series—I1

(Fig. 4—based on Iteration 1) Corruption (bci_bci) and GDP per Person (unna_gdppc) are the most important factors for prediction.

9 Deep learning setups

On a technical level, using and tuning even comparatively simple Feed Forward Neural Networks for prediction is in general more challenging and time consuming than applying supervised learning algorithms like k-NN and RF (Hastie et al. 2009: 397f.). A fact that should be considered by researcher planning to include networks in their research.

Feed Forward Neural Networks are an assembly of different layers of nodes. Nodes activate as soon as a certain threshold of input, a certain stimulus, was received from the previous layer. Data is send cascading through these layers and thus activates these nodes in characteristic patterns. Simplified speaking, the basic node activation process is in many ways comparable to how we imagine the human brain information processing, and thus is often called *neural* network (Goodfellow et al. 2016: 169). Patterns in the network are

formed and saved by regular activation of the involved neurons. Sending the training dataset through the network is thus repeated several times to adequately adapt the network to the data. Networks have a plentitude of parameters that must be configured and correctly set according to the applied data and the research question. Furthermore, the whole structure of the network must be designed, build and optimized accordingly. Depending on the specific network architecture and the applied parameters and optimizers it is not uncommon that a network might get stuck in a local minimum and entirely fails to adequately learn data, even though the technique is viable and would yield good results if applied correctly (Hastie et al. 2009: 400f). Judging the quality of this approach thus is trickier and more error prone, as failures to adequately build and optimize ANN structures might translate into sub-optimal predictions. ANNs do however achieve unmatched accuracy and predictive power in many advanced applications and should thus be included in this wide-ranging comparison of ML techniques.

For testing Neural Networks with the accumulated conflict dataset, the first series of tests involves a simple network pattern, that will be extended and fine-tuned in the following steps.

10 Single-hidden layer setup

The first series is conducted with a (feed forward) Network including a single hidden layer. Data is inserted in seven nodes resembling the seven variables of the dataset. From this input layer the data is passing through a hidden layer of 150 nodes and then passing on to the output layer with 5 nodes resembling the form of the categories which are to be predicted. The network setup is visualized in Fig. 5.

After testing this setup various times with different batch sizes and epochs, a batch size of 1 (so-called single or online processing) was chosen as the most fitting approach for the dataset as well as 250 epochs as sufficient amount of training (going beyond 250 epochs increasingly indicates reaching a local minimum and showing symptoms of overfitting). With 250 epochs the data passes 250 times the network overall to train and adjust it with its information. Batch size determines how many country-years are send trough before the network weights are adjusted to the new data. Higher batch sizes might be faster for processing, but lower batch sizes are more "adaptive", training the network in less epochs to the data. There are of course many more considerations about both parameters (Hastie et al. 2009: 397).

After each epoch, accuracy of prediction (lower frame) and loss (upper frame) on the training set (blue line) and a small (0.05–5%) validation set (green line) are printed (see Fig. 6).

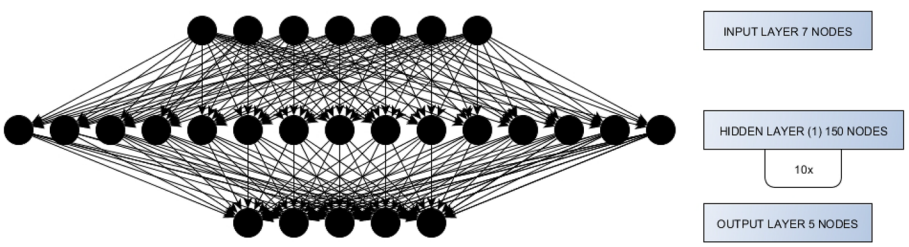

Fig. 5 Single hidden layer network setup

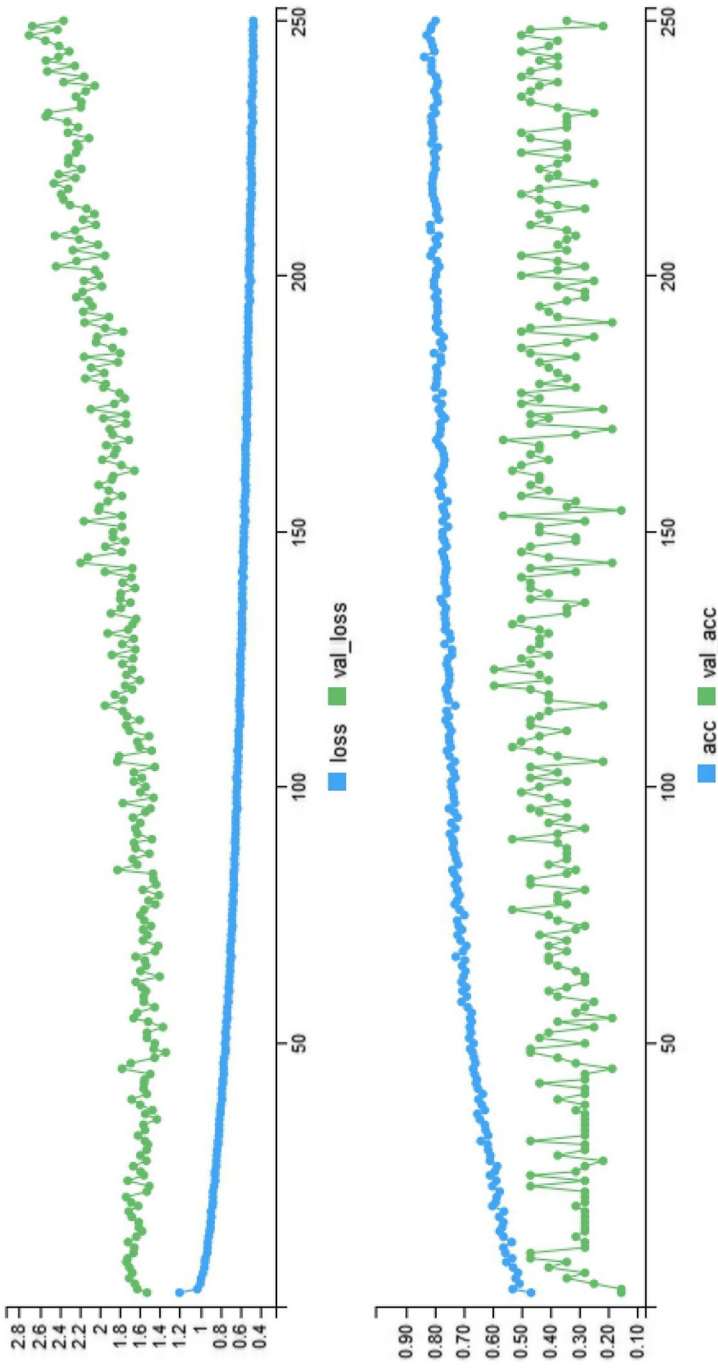

Fig. 6 Single hidden layer network prediction performance during learning process

Table 8 Single-Layer Feed Forward Network

5-cross validations *1-layer-250e–b1*	Accuracy on the test sample	Loss on the test sample
Iteration 1	0.6589862	0.8819193
Iteration 2	0.6632124	0.7567360
Iteration 3	0.6763285	0.9027525
Iteration 4	0.6714286	1.0099050
Iteration 5	0.6761905	0.9607828
Predict year data *1-layer-250e–b1*	Accuracy on the test sample	Loss on the test sample
Iteration 2013	0.6689655	1.0133940
Iteration 2014	0.6965517	0.8632736
Exclusion test *1-layer-250e–b1*	Accuracy on the test sample	Loss on the sample
Iteration A	0.4222222	2.292634
Iteration B	0.4647887	2.053121

Equally to the LDA, CART, k-NN and Random Forest Series, 641 observations are used in the training sample. Testing each series, a small number is excluded (5%) to form the validation sample, reporting after each epoch if unseen data is better identified or if overfitting occurs.

Because of the small training sample these excluded 5% of observations can already have a measurable impact on prediction quality. The network iterations thus have been performed once with and once without validation set.

The results reported below are without validation set to maintain the same amount of training cases in networks and supervised learning algorithms.

Like in the algorithm series, the network is used to predict the randomly selected 210 observations that were excluded before the training process started. Unsurprisingly, a basic network performs in its non-optimized state less accurate than the k-NN or Random Forest application for this task.

Table 8 shows the network performance in accuracy and loss on the test sample.

Surpassing linear regression models and CART by a good degree, this simple network still falls short of the predictive performance of the Random Forest Algorithm. It predicts approximately 11-percentage points less accurately than RF. This leads to roughly 3.3 times better predictions than simply guessing the category. Error on the Exclusion test is lower with 3-percentage points on the autocorrelation test iteration.

A general disadvantage of networks is the black box problem. In the previous paragraph it was shown, that it is easy to examine in RF setups, which variables had the biggest influence on the predictor consulting the Variance Importance Measures (VIMs). To find out which variables where most influential in networks is far more complicated. Leaving out variables and training separate iterations without certain elements is a time- and resource-consuming option. Predicting crafted test samples with maximum values for certain variables might be another option. Overall, as argued in the beginning, combining Random Forest or other suitable supervised learning algorithms systematically with neural network models might be a sensible solution in this aspect.

11 Triple hidden layer setup

The setup of the triple hidden layer setup is more complex and features dropout to improve the learning performance. The structure is displayed in Fig. 7.

The triple hidden layer setup includes two layers with dropout nodes. Each epoch a random composition of nodes is deactivated in both layers. In this example 20% of layer nodes deactivate each epoch, while 80% remain active. Dropout is a technique to effectively combat so called 'overfitting', a common problem with networks (Srivastava et al. 2014). Overfitting is synonymous to over adaptation to the training data. It leads to highly accurate predictions on the training sample, but if new test data is presented, the network fails to correctly handle and categorize this new unseen data. An overfitted network would perform well on data that is very similar composed as the training data, but would struggle with more diverse, divergent data and thus be very problematic to be accurately applied to a meaningful Conflict Prediction framework predicting on future developments.

Network performance in accuracy on the 3-layer test sample is presented in Table 9.

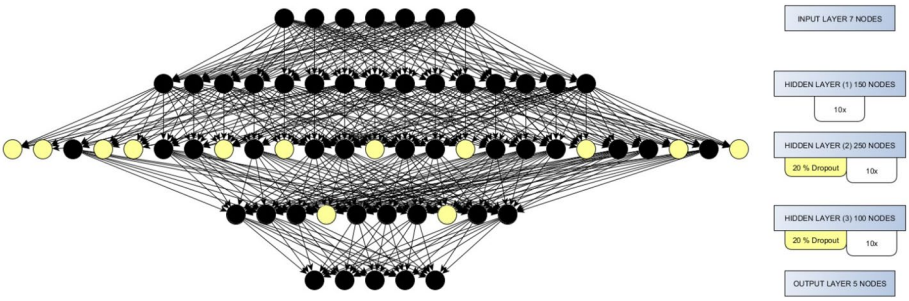

Fig. 7 Triple hidden layer architecture

Table 9 Triple hidden layer feed forward network with dropout (2 x 20%)

5-cross validations *3-layer-250e–b1*	Accuracy on the test sample	Loss on the test sample
Iteration 1	0.7419355	1.4177540
Iteration 2	0.7512953	0.9338002
Iteration 3	0.7487923	1.5240600
Iteration 4	0.7523810	1.3166150
Iteration 5	0.7333333	1.8810080
Predict year data *3-layer-250e–b1*	Accuracy on the test sample	Loss on the test sample
Iteration 2013	0.7379310	1.812775
Iteration 2014	0.7586207	1.466138
Exclusion test *3-layer-250e–b1*	Accuracy on the test sample	Loss on the sample
Iteration A	0.4311111	4.876879
Iteration B	0.4553991	4.308404

Table 10 Distribution of predictions in the triple-hidden layer prediction—iteration 2014

ANN *3 HL 2014*	Real				
Prediction	Peaceful	Almost peaceful	Minor conflict	Major conflict	War
Peaceful	**31**	2	0	0	0
Almost peaceful	7	**28**	5	0	0
Minor conflict	0	7	**26**	6	1
Major conflict	0	1	5	**21**	0
War	0	0	0	0	**5**

With 75% of accuracy this setup including two dropout layers provides a solid prediction and provides results that are only 3% points behind the RF prediction on the cross-validation and year-wise test samples. 5-CV results are 3.8 times better than randomly assigning observations to categories.

Noteworthy here is the relatively good performance improvement between the 2013 and 2014 prediction, indicating that the input of more data is comparatively more beneficial to the network than to some supervised learning models like k-NN or linear models like CART, that even suffer a decrease in predictive power by adding the additional year 2013 to the training data. On the other hand, the year 2014 has witnessed some impactful geopolitical changes (e.g. the Ukraine conflict) which may also contribute to the difficulties of predicting the new data year for simpler learning algorithms. Networks and the RF Algorithm managed to predict them better, non the less.

Table 10 indicates that the error of the triple hidden layer network with dropout is also strongly clustered around the correct categories. The network only predicted two country-years to be more than one category off from the real value. One "War" case was predicted as "Minor Conflict" and one "Almost Peaceful" case as "Major Conflict".

12 Recurrent LSTM-networks

Long-Short-Term-Memory Networks (LSTM) are together with other recurrent layer techniques like GRU or RELU on the forefront of new machine learning discoveries (Goodfellow et al. 2016: 410). Most sophisticated ML frameworks with big data foundations use them and achieve impressive results in tasks like image recognition (Kiros et al. 2014), speech learning (Graves 2013) or acquiring complex interactive behavior like playing the ancient game of Go at unparalleled levels (Silver et al. 2016).

LSTM Layer are fundamentally different compared to previously applied dense layers in Feed-Forward Networks, as every layer contains an internal structure, updating information in a much more complex, nonlinear way (Goodfellow et al. 2016: 408f). They are ideally fed with sequential data using the full potential of their internal updating mechanisms (Graves 2013; Sutskever et al. 2014). Recent works have proposed that RNNs can be successful applied to non-sequential data and still outperform normal Feed-Forward Networks by a significant degree (Chopra et al. 2017). Combining LSTM Layers with simple dense layers is the recommended baseline approach, using the following structure presented in Fig. 8.

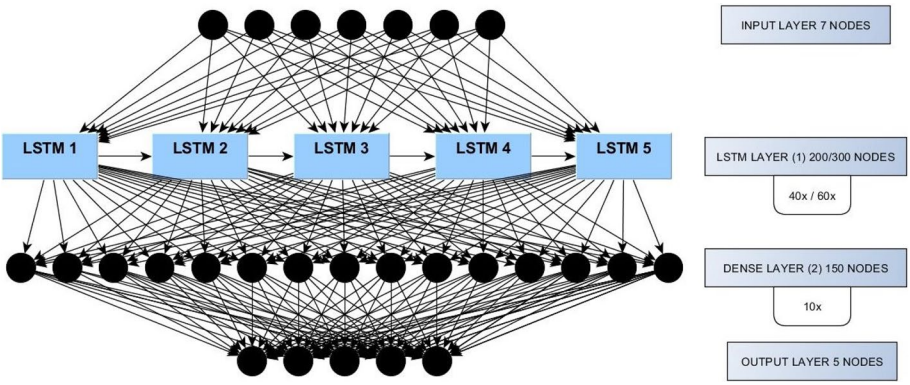

Fig. 8 LSTM network structure

Table 11 1-LSTM layer (200 nodes) + 1-dense layer—network

5-cross validations *1-lstm + dense-250e–b1*	Accuracy on the test sample	Loss on the test sample
Iteration 1	0.7714286	1.301360
Iteration 2	0.7761905	1.169862
Iteration 3	0.7761905	1.543415
Iteration 4	0.7571429	1.135024
Iteration 5	0.7619048	1.148386
Predict year data *1-lstm + dense-250e–b1*	Accuracy on the test sample	Loss on the test sample
Iteration 2013	0.7517241	1.554993
Iteration 2014	0.7655172	1.136749
Exclusion test *1-lstm + dense-250e–b1*	Accuracy on the test sample	Loss on the sample
Iteration A	0.3822222	4.121541
Iteration B	0.3286385	4.461779

One challenge in a LSTM framework is adapting the 2-dimensional Data to the 3-dimensional input layer. In my iteration, the simplest and most comparable approach to FFNNs is chosen, to transform the data table of dimensions 851×7 to a three-dimensional array maintaining these proportions with a single extra dimension of value 1 ($851 \times 7 \times 1$). As mentioned before, for using the full potential of LSTMs, restructuring the data to a time-series 3D array would be more beneficial for the learning process, but due to the limited size of each year's data and the overall dimensions of the dataset, such a remodeling of the data would not be sensible. Thus, any increase in predictive quality between FFNN and LSTM Networks solely results from the better learning and updating capacities of LSTM Layers and is not based on better structured data inputs.

LSTM Layers are excellent for complex problems, confronted with the rather simple data set, they can perform better than FFNNs, but even this advanced technique still

Table 12 1-LSTM layer (300 nodes) + 1-dense layer—network

5-cross validations *1-lstm300 + dense 200-250e*	Accuracy on the test sample	Loss on the test sample
Iteration 1	0.7857143	1.473617
Iteration 2	0.7714286	1.345390
Iteration 3	0.7857143	1.632399
Iteration 4	0.7857143	1.066597
Iteration 5	0.7904762	1.280574
Predict year data *1-lstm300 + dense 200-250e*	Accuracy on the test sample	Loss on the test sample
Iteration 2013	0.7448276	1.124691
Iteration 2014	0.8000000	1.246526
Exclusion test *1-lstm300 + dense 200-250e*	Accuracy on the test sample	Loss on the sample
Iteration A	0.3822222	3.748554
Iteration B	0.3568075	4.142637

performs not quite as good as supervised learning algorithms. Accuracy and Kappa of the LSTM Model are shown in Table 11.

Testing a Double-Layer LSTM Network, a setup effective in speech recognition, can perform marginally better on this data (< 1%).

Expanding the size of the single LSTM layer however does improve accuracy further and was chosen as the optimal way to improve the model further. Expanding the LSTM from 200 to 300 nodes increased the prediction quality by approximately 1.5% percentage points over all iterations. To prevent the network from overfitting and adapting to well to the training data, the number of iterations was automatically capped between 200 and 250 epochs, as soon as accuracy and loss increase fall below a certain threshold, a common approach in network setups.

The most efficient LSTM Network performance was assembled and is listed in Table 12.

With this architecture and layer size the network predicted the 2014 data equally well as a Random Forest Algorithm in terms of raw accuracy (80%) as visualized in Table 13. After testing over 40 network architectures, building and running iterations for hundredths

Table 13 LSTM 300n—prediction vs. real categories—iteration 2014

LSTM-300-*2014*	Real				
Prediction	Peaceful	Almost peaceful	Minor conflict	Major conflict	War
Peaceful	**34**	2	0	1	0
Almost peaceful	4	**30**	4	0	0
Minor conflict	0	6	**26**	5	1
Major conflict	0	0	5	**21**	0
War	0	0	1	0	**5**

of hours, including and testing LSTM setups in conflict prediction for the first time, one network setup performed on par with the random forest algorithm, which took a fraction of the time to build and almost no effort to configure. Interestingly, the 20% of incorrectly assigned cases are very differently distributed over the 5-Category table compared to the RF results. As networks and algorithms learn patterns in a completely different way, they also err differently. This is good for prediction frameworks, as network and RF results can be systematically compared and cross-checked. Cases that are equally predicted by several approaches could be considered safer, more solid predictions. Cases with dissimilar prediction results should be carefully observed and evaluated. This comparability is another solid argument for applying Neural Networks and Supervised learning algorithms as often as possible together in upcoming research designs.

13 Comparison summary of all prediction techniques

In the following graphs all introduced methods and architectures are compared by accuracy on the test sample, including all selected random test samples as well as the year 2013/2014 test samples. The results of several additional supervised learning algorithms are added, specifically Support Vector Machines, Bagged Tree models and C5.0 are added to compliment the comparison of supervised ML techniques. Figure 9 shows a comparison of all techniques in order of appearance. Figure 10 provides the year 2013 and year 2014 predictions ranked by best accuracy.

From this comprehensive comparison, a list of observations regarding the accuracy of the predictions can be drawn: Both networks and supervised learning algorithms outperform linear and regression-based predictions reliably and consistently by a wide margin on the small-to-medium sized data set.

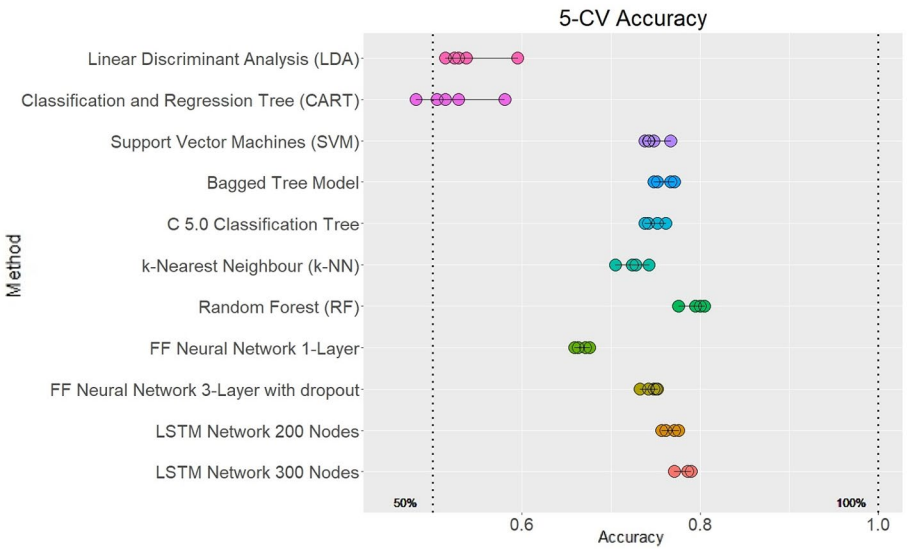

Fig. 9 Comparison of all fivefold cross-validations, in order of appearance

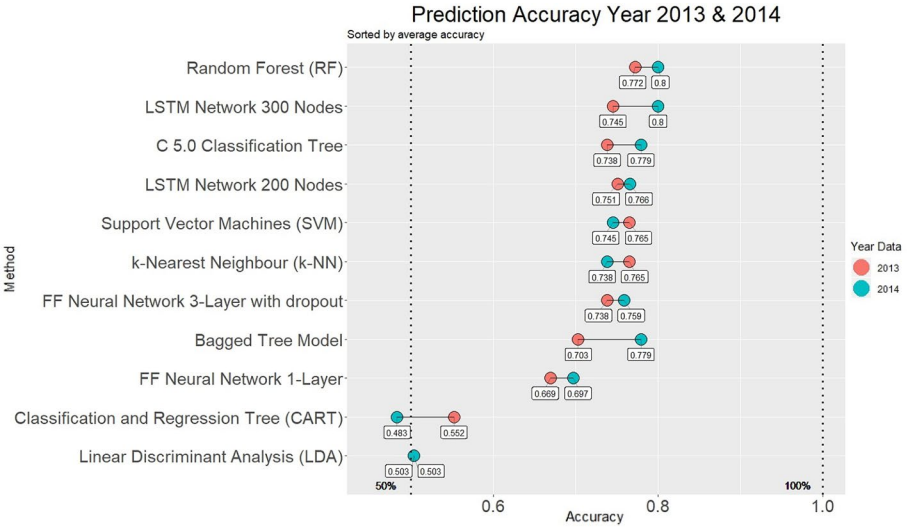

Fig. 10 Comparison of the year 2013 and 2014 predictions, in order of accuracy

Networks occasionally outperform supervised learning algorithms, but not consistently. Network performance in conflict prediction depends crucially on the setting, architecture and hyperparameter tuning. Some network architectures achieve higher prediction values; some only surpass linear predictive models reliably.

LSTM-Layer Network Models work well for predicting conflict levels in this conflict data setting, despite them being predominantly applied and associated to speech or image recognition in sequential data settings. As LSTM-Layers have, up to the current date, never been applied and thoroughly tested in conflict prediction before, and only occasionally in non-sequential settings overall, this is a highly relevant and interesting result for future conflict prediction frameworks.

Another interesting observation from comparing supervised learning algorithms and deep learning networks is, that the error shown in the cross-tables (Tables 7, 10 and 13) is systematically different between a variety of algorithms and networks. The Year 2014 iteration results reveal that the RF Algorithm falsely assigned two country-years of conflict category "Major Conflict" to "Almost Peaceful", committing the strongest classification deviation from the real categories. The two falsely assigned cases in this iteration are the country years "El Salvador 2014" and "The Philippines 2014". "El Salvador 2014" with an GPI Score of 2.263 and "The Philippines 2014" with a score of 2.462 are both located relatively close to the border of the two categories "Minor Conflict" and "Major Conflict" (separation at 2.241). Other years of the country El Salvador, for example 2013 and 2010 therefore were classified as "Minor Conflict". Furthermore, the Philippines and El Salvador are very specific cases. El Salvador's high rating as a

major conflict is largely related to organized crime and gang violence. The Philippines even higher rating is heavily related to terrorism and the escalating violence in Mindanao Province.

The LSTM Network strong errors varied from the Random Forest Patterns. A case that was in various iterations assigned to wrong categories was "India 2014", often assigned as "Almost Peaceful", although the real GPI score of 2014 indicated "Major Conflict". Again in the case of India regional flash points for example in Kashmir and localized terrorist attacks from Maoist and nationalistic groups lead to significant higher conflict rating while the socio-economic, political and demographic indicators of the whole country might indicate lower levels and lead to lower predictions. Despite these miss-classifications of complicated, multilayers cases with different conflict settings like organized crime, terrorism, civil war and/or interstate war, the applied predictive frameworks managed consistently to assign more than 80% correctly.

Using both techniques, Algorithms and Networks together, controlling for identical and deviating predictions is a viable way to identify atypical cases and to validate and double-check prediction results. This can further boost framework performances.

The final results of this categorical approach can also be compared to regression-based approaches and random forest regression tree results obtained with a continuous conflict variable (the original GPI Score without classification in categories). A detailed excursion on predicting on a continuous variable with random forest can be found in the "Appendix".

Additionally, in the "Appendix" the problems of missing data and autocorrelation is discussed in detail.

Since only 7 variables and 851 cases were included in this framework, it is very well possible that Random Forest and LSTM Networks already achieved a near-optimal prediction state for this small setup due to the limitations of empirical information. To increase prediction quality beyond 80% simply a more elaborate and extensive data foundation might be necessary, as conflict escalation is a complex, multilayered process. Based on the simple data framework presented in this paper, variables and additional years can now be added step-by-step and their impact can be assessed after each update, building the foundation for an organically improving setup.

14 Conclusion

Cederman and Weidmann noted in their article about the future of conflict prediction that "to assess the added value of new approaches, analysts need to do a better job comparing their forecasts from complex prediction machinery to simple baseline models." (Cederman and Weidmann 2017: 476) This paper tries to follow exactly this recommendation by systematically comparing different networks and supervised learning methods with simpler linear and regression approaches on the same data set and different splitting methods.

Using this conflict data set as an example, a series of discoveries regarding the implementation of deep learning and machine learning algorithms in conflict prediction are highlighted. Especially, that using annual GPI conflict data is a viable target variable for a successful conflict predictor framework.

Two potential concerns with the presented target variable have been mentioned in this paper:

The Global Peace Index as a data foundation for conflicts might integrate non-conflict data due to the ambiguity in the GPI Scoring system. As argued before, this effect has been tested and found to be small enough to go without time-intense reconfiguration of the conflict data as a categorical system is formed on top of this data.

For the second point, the mixing of inter- and inner-state conflict events might be a problem if a deeper understanding of each phenomenon is the overall objective. Yet, from a prediction perspective, the objective of predicting conflict intensity in a country does not require to build separate frameworks for each specific type of conflict. By mixing all conflict types together the algorithm or networks, not the researcher, is required to learn how to identify specific symptoms within the economic indicators pointing to inter-state or inner-state or any hybrid form of conflict.

On the side of methodological contributions, this paper adds new observations regarding networks in small conflict prediction setups:

First and foremost, modern LSTM Networks are tested for conflict prediction and successfully applied to the described data framework. Using the LSTM Architecture with two-dimensional data still provides an improvement of accuracy, despite the predominant association of this technique to 3-D sequential data.

Networks do however not automatically outperform supervised learning algorithms in conflict forecast settings on a country level, as they are presumably requiring more data inputs to be effectively better then supervised algorithms. They are capable of reaching the predictive power of the most efficient and well-tested supervised learning algorithms and might be highly useful to include in prediction frameworks to compare results and affirm or contest predictions. After running over 40 different network architectures including complex LSTM layers and fine tuning over 25 independent hyperparameters, the best network iterations reached a Random Forest Algorithm in the accuracy of identifying the conflict levels of country-years.

In the “Iteration 2013”, only by training the indicator patterns available between 2009 and 2012, predicting on the socio-economic and political data assembled of 2013, the conflict situation in the next 12 months can be predicted with high accuracy by various techniques. RF and advanced Networks even performed better in the “Iteration 2014”, learning all country-years from 2009 till 2013 to predict the conflict situation of 2015 based on 2014 socio-economic and political data. They both achieve accuracies of up to 80% in the corresponding iterations. The predictions of 2013 and 2014 cases and the applied k-fold cross-validation results are almost identical in accuracy, this indicates that random sampling of k-fold CV can be assumed to be a reliable general benchmark for prediction quality in future years included in the setup.

However, with the Exclusion Data frame split by countries, separating data from the learning process and predicting years of countries completely unseen to the network, the LSTM networks did experience a heavier drop in prediction accuracy than networks with drop-out layers. Their sensitivity to abstract data can be compared to the drop of accuracy in the k-NN Algorithm series during the Exclusion Test. This might indicate that Networks with several dropout layers could be better suited for more abstract data than a pure LSTM setup. Combining LSTM with dense drop-out layers could be solution in this regard and is target for further research.

Regarding the applicability as interdisciplinary research method in political conflict research, it can easily be confirmed in this paper, that networks are several times more complex to configure and adjust to the research question and that they require a lot of mathematical and computational expertise to establish. They also return only limited feedback on how they predict and weight information.

If socio-economic data from 2016 to 2018 is rapidly incorporated, the prediction quality might not only increase further but also viable real-time predictions on upcoming years might be possible with this up-scalable framework soon. Alternatively, the predictor can always use simulated or forecast data to make a viable prediction about future conflict development scenarios.

Based on this research, as an illustrative example, I propose that networks, despite being excellent in many big data learning environments and despite all the justifiable hype (Hastie et al. 2009: 392) associated with them, should not necessarily replace simpler machine learning techniques in many common research settings featuring small and medium data sets, as mostly prevalent in political science. Networks can be a powerful complement for existing techniques and to diversify a framework. Overall, depending on the heterogeneity and cohesion of the data a supervised learning algorithm might still reliably outperform most network models even after extensive fine-tuning and building and testing dozens of model architectures.

Especially if working with a limited number of variables and cases or limited time and resources, networks might not be recommended as exclusive approach. Especially, if understanding the prediction process and evaluating the impact of all variables is part of the objective, networks provide very little added value as singular interdisciplinary approach. Applying networks together with supervised learning algorithms seems sensible to control for and mitigate many adverse tuning effects, the black box problem, and to control the efficiency and predictions of the applied network architecture.

Acknowledgements The author would like to thank the anonymous reviewers for the valuable comments and discussion.

Appendix

Random forest compared to (mixed effects multi-level) regression

To compare a maximum of available machine learning methods in this paper a categorical system was chosen to be put on top of the dependent variable. If a deeper comparison

Table 14 Regression results: OLS

Source	SS	df	MS		Number of obs = 851	
					F(7, 843) = 246.68	
Model	115.92847	7	16.5612101		Prob > F = 0.0000	
Residual	56.5953236	843	.067135615		R-squared = 0.6720	
					Adj R-squared = 0.6692	
Total	172.523794	850	.202969169		Root MSE = .25911	
gpi_ny2	Coef.	Std. Err.	t	P > |t|	[95% Conf. Interval]	
bci_bci	.0110257	.001176	9.38	0.000	.0087175	.0133338
fh_ep	.0012389	.0044535	0.28	0.781	− .0075025	.0099802
ffp_ued	.0253074	.0088165	2.87	0.004	.0080024	.0426123
dr_eg	− .0002198	.000837	− 0.26	0.793	− .0018626	.0014231
unna_gdppc	2.88e − 06	8.57e − 07	3.36	0.001	1.19e − 06	4.56e − 06
fh_fotpsc	.0050597	.0009054	5.59	0.000	.0032827	.0068367
ffp_ref	.0730338	.0065697	11.12	0.000	.0601388	.0859287
_cons	.6428183	.138931	4.63	0.000	.370127	.9155096

between machine learning and different regression methods is intended, the categorization of the conflict variables is not necessary or helpful.

To complete the comparison of different predictive models a random forest regression tree model (RF-RTM), a special for of random forest algorithm, is compared in the following part with several regression models including a nested multi-level model.

Different regression models achieve with the provided data between 64 and 67% of explained variance on the dataset. A linear model based on the metric conflict variable is shown in Table 14. Table 15 contains a multi-level mixed effects Model separated by countries.

A random forest prediction over the full dataset (no training and testing set, but the full exclusion cases with missing data) provides the following results demonstrated in Table 16 of the "Appendix".

Table 15 Regression Results: Mixed-effects ML regression

Mixed-effects ML regression				Number of obs = 851		
Group variable: country				Number of groups = 144		
				Obs per group		
				min = 1		
				avg = 5.9		
				max = 18		
				Wald chi2(7) = 391.42		
Log likelihood = 431.54096				Prob > chi2 = 0.0000		
gpi_ny2	Coef.	Std. Err.	z	P > |z|	[95% Conf. Interval]	
bci_bci	.0160845	.0018089	8.89	0.000	.0125392	.0196298
fh_ep	− .0048588	.0044866	− 1.08	0.279	− .0136523	.0039347
ffp_ued	.0259111	.0088171	2.94 0.003	.0086299	.0431922	
dr_eg	− .0009582	.0011394	− 0.84	0.400	− .0031914	.0012751
unna_gdppc	2.97e − 06	1.17e − 06	2.54	0.011	6.76e − 07	5.27e − 06
fh_fotpsc	.0010418	.0009622	1.08	0.279	− .0008441	0029277
ffp_ref	.0613059	.0079864	7.68	0.000	.0456528	.076959
_cons	.729326	.1808786	4.03	0.000	.3748103	1.083842
Random-effects parameters		Estimate	Std. Err.		[95% Conf. Interval]	
country: Identity						
	var(_cons)	.059217	.0074879		.0462182	.0758717
	var(Residual)	.0119623	.0006409		.01077	.0132867
LR test vs. linear model: chibar2(01) = 971.49					Prob > = chibar2 = 0.0000	

Table 16 Random forest regression tree results

Sample size of test (predict) data: 851
Number of grow trees: 500
Average no. of grow terminal nodes: 169.634
Total no. of grow variables: 7
Resampling used to grow trees: swr
Resample size used to grow trees: 851
Analysis: RF-R
Family: regression tree
Percentage (%) of variance explained: 88.47%
Test set error rate: 0.12

Variable importance measures (full set)

bci_bci	fh_ep	ffp_ued	dr_eg	unna_gdppc	fh_fotpsc	ffp_ref
0.2847696	0.2147070	0.2344239	0.1367566	0.1892967	0.1882165	0.3796357

Random forest provides us with up to 88.47% of explained variance on the full set with the metric conflict data. Linear regression models and mixed effects multi-level regression models achieve a maximum of up to 67% of explained variance. As shown in Fig. 11, the relationship between the socio-economic, political and demographic

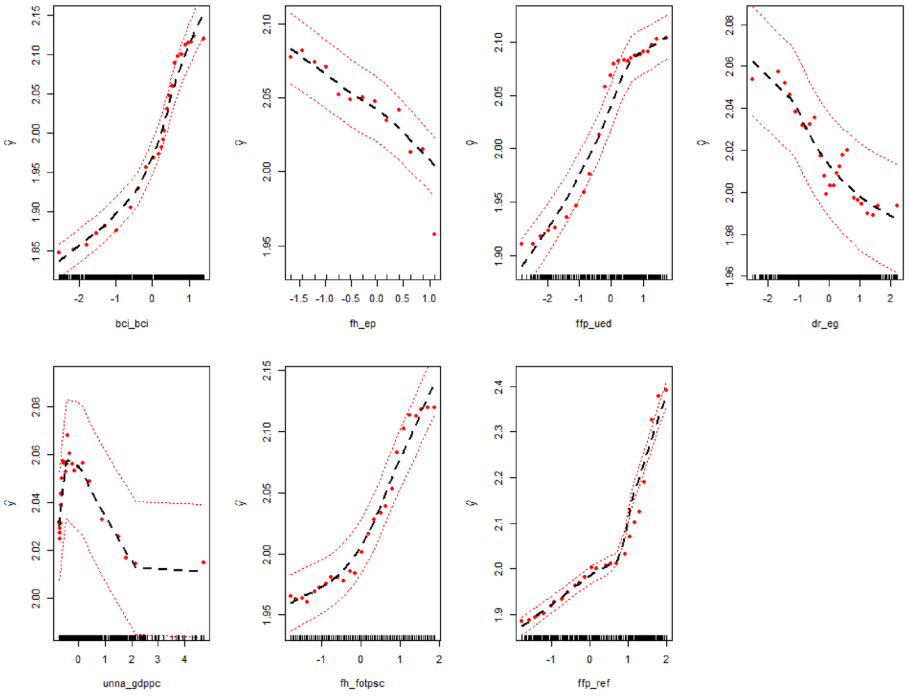

Fig. 11 Non-linear relationship between predictor variables and GPI score (Effect of variables on the GPI score)

variables and the dependent variable (GPI) is clearly non-linear. Thus, Random Forests improves the prediction of the non-categorical model by over 20%. This indicates a huge advantage of ML methods over Regression-based analysis with metric conflict data expressions, in accordance with what we observed in the categorical iterations in the main part of the paper.

Imputation of missing values

A very important issue in conflict research and prediction is how to handle missing data. As missing values occur more frequently in cases with higher conflict levels, excluding all cases with missing values would systematically exclude many high conflict intensity cases. In the previous parts of this publication, only complete cases were used to provide a maximum of comparability between the eleven different prediction methods.

Many machine learning techniques, including random forest, can handle missing values in efficient ways: One simple way is imputation by average (Breiman 2001), a better way is imputation by using a MissForest Model (Tang and Ishwaran 2017; Stekhoven and Bühlmann 2011) first to predict missing values according to the learned neighborhood of these cases. Using the RF imputation technique, cases with missing values can be imputated based on their similarity to cases with available data.

To demonstrate this technique a RF model is generated to predict all 924 cases that can be completed by imputation to predict the overall conflict intensity. 73 cases can be included and predicted in this model, which previously had to be excluded in the main prediction series. Ten of these country-year cases with incomplete data from the dataset are presented in Table 17 and compared to their real GPI scores to show the potential of this imputation method.

This advanced imputation method can be used combined with all eleven ML techniques presented in the main paragraphs. In the main part of the paper this technique was not applied in order to not further increase the complexity of the contribution.

Autocorrelation

Autocorrelation cannot be entirely avoided with the presented framework. Imitating a multi-level approach like in regression based models would reduce the amount of learnable data to a very small sub-sample, which is unfit for systematically learning patterns with algorithms. But due to the multinomial nature of ML predictions (Shown in Fig. 11, "Appendix"), it can be argued that mixed effects multi-level designs are not as beneficial to machine learning predictions as to linear models, as ML "curves" for prediction are resembling higher polynomial curves which automatically fit better on top of fractured, multi-year data. Furthermore, this paper argues that loosing multi-level applicability is more than compensated by the more than 20% higher baseline predictive accuracy of ML techniques.

Still, the sensitivity to changes of singular cases can be tested: If autocorrelation is expected to be extremely strong, the categorization of cases should not easily change

Table 17 Prediction of 10 additional country-years with incomplete data

Country-year	*bci_bci* Corruption	*fh_ep* Electoral process	*ffp_ued* Uneven economic development	*dr_eg* Globalized economy	*unna_gdppc* GDP per capita	*fh_fotpsc* Freedom of the press	*ffp_ref* Refugee pressure	*gpi_ny* Real GPI score	Predicted GPI score
Cuba 2013	40,6372	0	5,9	**NA**	6789,88	90	5	2002	2033
Djibouti 2011	51,3585	3	7,145209	**NA**	1372,37	74	6,9	1933	2110
Iraq 2010	62,8400	7	8,998107	**NA**	3794,78	68	9	3342	2734
Somalia 2012	70,6729	0	8,4	**NA**	130,15	84	10	3394	3005
South Sudan 2012	59,8863	4	8,934884	**NA**	944,28	60	10	2602	2789
Uzbekistan 2014	62,0702	0	7	**NA**	2138,78	95	6	2187	2420
Yemen 2014	**NA**	3	8,1	50,43277	1418,08	78	9,1	2751	2791
Sudan 2011	65,2156	2	8,808772	33,46134	**NA**	78	9,9	3398	3132
Liberia 2009	50,1776	8	8,3	**NA**	267,96	61	8,2	2148	2230
Israel 2013	36,6590	12	**NA**	73,67835	37403,24	30	**NA**	2725	1738

Table 18 China 2035: Chinas economic rise defies democratization

cname_year	bci_bci	fh_ep	ffp_ued	dr_eg	unna_gdppc	fh_fotpsc	ffp_ref	gpi_ny
China 2014	47,645672	0	7,4	49,9711952	7616,70655	86	5,6	2267
China 2035	40	0	7,2	62	26000	88	5,6	~ 1,85–2,25

Development scenario: No democratic reforms and continued strong crackdown on media, moderate corruption reduction, considerable increase in GDP

Predicted conflict intensity: **Minor conflict**

with one or two variables deviating while largely maintaining the pattern of the remaining 5–6 parameters anchored to the specific country case.

25 plausibility tests were conducted with Random Forest, involving country-years with slight modifications to their variable sets. These tests evaluate under which conditions these cases switch their predicted conflict intensity category. Three results are presented in so-called "Conflict Scenarios" for a fictive year 2035 in the following paragraph. The results indicate that Country years do switch at sensible thresholds if one or only two parameters change.

The algorithm proposes that China (Table 18) might be able to preserve its authoritarian regime type without risking higher conflict levels in the long run.

The algorithm results also propone that Venezuela (Table 19) might achieve a more peaceful state than an autocratic China by 2035, but only if it becomes a fully open democratic system with improving economic integration in the global economy and with extensive freedoms. An economic sunshine and liberalization scenario after President Maduro, so the implication, would come with a strong peace dividend.

A more democratic and liberal but economic still sluggish Nigeria (Table 20) suffering perpetual corruption and highly uneven development would foster a better outlook

Table 19 Venezuela 2035: Full democratization and economic recovery

cname_year	bci_bci	fh_ep	ffp_ued	dr_eg	unna_gdppc	fh_fotpsc	ffp_ref	gpi_ny
Venezuela 2014	68,306488	5	6,7	43,2668877	16614,6833	81	4,8	2493
Venezuela 2035	60	11	5,5	55	24000	30	4,6	~ 1,6–1,85

Development scenario: Slight reduction of corruption, substantial democratization, less uneven development, a more globalized economy, higher GDP, consolidated press freedoms

Predicted conflict intensity: **Almost peaceful**

Table 20 Nigeria 2035: More liberties but sluggish economy and rampant corruption

cname_year	bci_bci	fh_ep	ffp_ued	dr_eg	unna_gdppc	fh_fotpsc	ffp_ref	gpi_ny
Nigeria 2014	65,583038	6	8,8	55,0973511	3203,24344	53	7,5	2,91
Nigeria 2035	68	8	8,8	56	4000	78	7	~ 1,85–2,25

Development scenario: Slightly more corruption (from a very high starting point), improved electoral process, slightly higher GDP (from a low basis), significant improvements to press freedoms

Predicted conflict intensity: **Minor conflict**

than in the current state. Although these tests are merely anecdotal evidence, they indicate that the prediction framework is sensible to singular changes and would thus be capable of correctly predicting conflict intensity after radical effects like regime change, economic collapse or mass migration events.

References

Alpaydin, E.: Introduction to Machine Learning. MIT Press, Cambridge (2004)

Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Comput. **9**, 1545–1588 (1997)

Beck, N., King, G., Zeng, L.: Improving quantitative studies of international conflict: a conjecture. Am. Political Sci. Rev. **94**, 21–35 (2000)

Benitez, J.M., Castro, J.L., Requena, I.: Are artificial neural networks black boxes? IEEE Trans. Neural Netw. **8**, 1156–1164 (1997)

Bishop, C.M.: Pattern recognition and machine learning. Springer, Singapore (2006)

Breiman, L.: Bagging predictors. Mach. Learn. **26**, 123–140 (1996)

Breiman, L.: Heuristics of instability in model selection (Technical Report). University of California at Berkeley, Statistics Department, Berkeley (1994)

Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)

Caruana, R., Niculescu-Mizil, A., 2006. An Empirical Comparison of Supervised Learning Algorithms, in: Proceedings of the 23rd International Conference on Machine Learning. Presented at the ACM, Pittsburgh

Cederman, L.-E., Weidmann, N.B.: Predicting armed conflict: time to adjust our expectations? Science **355**, 474–476 (2017)

Chadefaux, T.: Conflict forecasting and its limits. Data Science **1**, 7–17 (2017)

Chopra, C., Sinha, S., Jaroli, S., Shukla, A., Maheshwari, S.: Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients. In: Proceedings of ICCBB. Presented at the International Conference on Computational Biology and Bioinformatics. pp. 18–23. Newark, NJ (2017)

Colaresi, M., Mahmood, Z.: Do the robot: lessons from machine learning to improve conflict forecasting. J. Peace Res. **54**, 193–214 (2017)

Collier, P., Hoeffler, A.: On the economic causes of civil war. Oxford Econ. Papers **50**, 563–573 (1998)

Collier, P., Hoeffler, A.: Greed and grievance in civil war. Oxford Econ. Papers **56**, 563–595 (2004)

Collier, P., Hoeffler, A.: Resource rents, governance, and conflict. J. Conflict Resolut. **49**, 625–633 (2005)

Cramer, C.: Does inequality cause conflict? J. Int. Dev. **15**, 397–412 (2003)

Cutler, A., Cutler, R., Stevens, J.R.: Random Forests. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning, pp. 157–175. Springer, Boston (2012)

Dorussen, H.: Balance of power revisited: a multi-country model of trade and conflict. J. Peace Res. **36**, 443–462 (1999)

Dreher, A.: Does globalization affect growth? Evidence from a new index of globalization. Appl. Econ. **38**, 1091–1110 (2006)

Ek, R., Karadawi, A.: Implications of refugee flows on political stability in the Sudan. Environ. Secur **20**, 196–203 (1991)

Fearon, J.D., Laitin, D.D.: Ethnicity, insurgency, and civil war. Am. Political Sci. Rev. **97**, 75–90 (2003)

Feder, D.R.: Political instability, refugees, and a health care crisis in Malawi. East Afr. Geogr. Rev. **20**, 47–57 (1998)

FFP: Fragile States Index and Cast Framework Methodology (Technical Report). Fund for Peace, Washington, D.C. (2017)

Gohdes, A.R., Carey, S.C.: Canaries in a coal-mine? What the killings of journalists tell us about future repression. J. Peace Res. **54**, 157–174 (2017)

Goldstone, J.A., Bates, R.H., Epstein, D.L., Gurr, T.R., Lustik, M.B., Marshall, M.G., Ulfelder, J., Woodward, M.: A global model for forecasting political instability. Am. J. Political Sci. **54**, 190–208 (2010)

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

Graves, A.: Generating sequences with recurrent neural networks. ArXiv Technical Report (2013). arXiv :1308.0850

Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
Hegre, H.: Trade decreases conflict more in multi-actor systems: a comment on Dorussen. J. Peace Res. **39**, 109–114 (2002)
Hegre, H., Buhaug, H., Calvin, K., Nordkvelle, J., Waldhoff, S.T., Gilmore, E.: Forecasting civil conflict along the shared socioeconomic pathways. Environ. Res. Lett. **11**(5), 054002 (2016)
Hegre, H., Sambanis, N.: Sensitivity analysis of empirical results on civil war onset. J. Conflict Resolut. **50**, 508–535 (2006)
Ho, T.K.: Random Decision Forests. AT&T Bell Laboratories, Murray Hill (1995)
Hudson, V.M., Schrodt, P., Whitmer, R.: Discrete sequence rule models as a social science methodology: an exploratory analysis of foreign policy rule enactment within Palestinian-Israeli event data. Foreign Policy Anal. **4**, 105–126 (2008)
King, G., Zeng, L.: Logistic regression in rare events data. Political Anal. **9**(2), 137–163 (2001)
Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models (2014). ArXiv arXiv:1411.2539
Le Billon, P.: Buying peace or fuelling war: the role of corruption in armed conflicts. J. Int. Dev. **15**, 413–426 (2003)
Muchlinski, D., Siroky, D., He, J., Kocher, M.: Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Anal. **24**, 87–103 (2016)
Neudorfer, N.S., Theuerkauf, U.G.: Buying war not peace: the influence of corruption on the risk of ethnic war. Comp. Political. Stud. **47**, 1856–1886 (2014)
O'Brien, S.P.: Crisis early warning and decision support: contemporary approaches and thoughts on future research. Int. Stud. Rev. **12**, 87–104 (2010)
Perry, C.: Machine learning and conflict prediction: a use case. Int. J. Secur. Dev. **2**, 1–18 (2013)
Sarkees, M.R., Wayman, F.W., Singer, J.D.: Inter-state, intra-state, and extra-state wars: a comprehensive look at their distribution over time, 1816–1997. Int. Stud. Q. **47**, 49–70 (2003)
Schmeidl, S., Bond, D.: FAST Early Warning and Conflict Prevention: The Strengths and Limitations of (Automated) Event-Data Monitoring to Support Early Warning Analyses. Paper prepared for the annual convention of the International Studies Association. Swiss Peace Foundation, Berne (2000)
Segal, M.R.: Machine Learning Benchmarks and Random Forest Regression (Technical Report). Center for Bioinformatics & Molecular Biostatistics, San Francisco (2003)
Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., von der Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M.: Mastering the game of Go with deep neural networks and tree search. Nature **529**, 484–489 (2016)
Smidt, M., Vernaccini, L., Hachemer, P., De Groeve, T.: The Global Conflict Risk Index (GCRI)—Manual for data management and product output (No. Version 5). GCRI (2016)
Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
Standaert, S.: Divining the level of corruption: a Bayesian state-space approach. J. Comp. Econ. **43**, 782–803 (2015)
Stekhoven, D.J., Bühlmann, P.: MissForest—nonparametric missing value imputation for mixed-type data (2011). ArXiv arXiv:1105.0828v2 [stat.AP]
Stewart, E. (ed.): Horizontal Inequality and Conflict. Palgrave, New York (2008)
Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks (2014). ArXiv arXiv:1409.3215 [cs.CL]
Tang, F., Ishwaran, H.: Random Forest Missing Data Algorithms (2017). ArXiv arXiv:1701.05305v2 [stat.ML]
Vreeland, J.R.: The effect of political regime on civil war: unpacking anocracy. J. Conflict Resolut. **52**, 401–425 (2008)
Ward, M.D., Greenhill, B., Bakke, K.: The perils of policy by p-value: predicting civil conflicts. J. Peace Res. **47**, 363–375 (2010)