Electoral Vulnerability Index 2024
Technical report, September 2024

David Randahl[1,2], Maxine Leis[1,2], and Tim Gåsste[1,2]

[1] *Department of Peace and Conflict Research, Uppsala university*
[2] *UpSight QRA*

1 Introduction

During the spring 2024 the UpSight team has on the behalf of the Kofi Annan Foundation (KAF) worked to update the Electoral Vulnerability Index (EVI), a forecasting tool for forecasting electoral violence. The EVI was originally developed in 2022 using state of the art machine learning techniques. The system uses data from the Varieties of Democracy (V-DEM) project (Coppedge et al., 2021b), as well as data from the Digital Societies Project (DSP) (Mechkova et al., 2022), and the World Bank's World Development Indicators (WDI). Forecasts are made annually for each country with at least one scheduled election in the specific year, and made for two calendar years into the future from the last available data.

This technical report briefly outlines the definitions and methodology used by the EVI and highlights some changes in the methodology of the EVI since the first EVI report. This report also shows the updated evaluation metrics of the tool and the forecasts for 2024 and 2025. The report will be followed by an academic article outlining the project in further detail.

2 Defining electoral violence

The first task of the D-Arch team was to develop an outcome measure for Electoral Violence which is conceptually sound and with properties which makes it practically useful for the KAF and which makes it possible to properly forecast. In discussions with KAF about what types of forecasts are most useful in practice, and internal discussions about what data are possible to use, we landed in agreeing that an ordinal outcome for electoral violence was the most useful. Based on a review of the existing sources of electoral violence in Fjelde et al. (2021) we decided to use the Varieties of democracy project indicators of electoral violence. To code this outcome we used two indicators measuring electoral violence by non-government actors (v2elpeace) and intimidation and harassment by government actors (v2elintim) from the VDEM election-level data. These two indicators are themselves ordinal on a 5-point scale from 0 (most electoral violence/harassment) to 4 (least electoral violence/harassment). These indicators are coded by country experts, and we use the mean value across all coded values (Coppedge et al., 2021a). The full description of these two indicators and their levels can be found in Appendix A.

In order to make the forecasts more stable and to make a more conceptually clear outcome, we re-coded these two indicators into a three-point ordinal scale measuring 'electoral violence and/or harassment'. We did this by first re-coding the 5 point scales to a three point ordinal scales by merging the categories by coding values belolw 1.5 as 'severe electoral violence or harassment', values between 1.5 and 3 as 'limited electoral violence or harassment' and values of 3 or above as 'no electoral violence or harassment'. We then coded our outcome, 'electoral violence and/or harassment', as the most severe value across these two three-point indicators. This definition differs slightly from the definition used in the 2022 version of the EVI where the category of 'no electoral violence' was coded for elections with a value of 3.5 or above on the 5 point scales. This adjustment was made to be a bit more lenient in the coding of peaceful elections and thus to more effectively separate between the no violence and limited violence categories. Additional reflections on how this change affects the forecasts are outlined in the section on changes since the last EVI below.

One important caveat with regards to this outcome measure is that this indicator captures *both* electoral violence *and* intimidation/harassment by the government. One consequence of this is that elections where the government repression has been so severe as to cause an artificially calm election with no outbursts of visible violence are also coded as having 'severe' levels of electoral violence since the government repression is so severe. Another implication of this definition is that countries with flawed elections are almost certain to experience electoral violence and thus yielding high levels of forecasted risk. To account for this we present the true forecasts separately for countries deemed by the VDEM institute to be democracies separately from those deemed to be autocracies.

One alternative way of dealing with this potential problem is to separate the forecast into government harassment and/or intimidation as one category and electoral violence by other actors as a second category. Each country would, in that case, get a separate forecast for each of these two outcomes. This would make communication about the outcome more complex, but perhaps more in line with how electoral violence is conceptualized outwardly.

2.1 Forecasting the outcome

Since the outcome of electoral violence is ordinal the forecasts for the outcome show the probability that each election end up in the three categories of 'no electoral violence/harassment', 'limited electoral violence/harassment', and 'severe electoral violence/harassment'. In order to facilitate an easier interpretation of the results, two additional measures are also presented. First, the likelihood that *any* electoral violence occurs for the election, i.e. simply the sum of the probabilities of 'limited' and 'severe' electoral violence/harassment. Second, we also present a 'risk index' scaled from 0 to 1 where 0 indicates the least risk of electoral violence and 1 the highest risk of electoral violence. The risk index in constructed by simply taking the probability of 'severe' electoral violence and adding 0.5 x the probability of 'limited' electoral violence. We make these forecasts up to two years into the future.

3 Data and methodology

The forecasting methodology used for the project is anchored in the methodologies used by the Violence Early Warning System (ViEWS) to predict violence from armed conflict (Hegre et al., 2019; Hegre et al., 2021). To this end, train machine learning models on historical data with features (predictor variables) grouped into broad thematic *constituent models*. The predictions of these constituent models are then combined into an *ensemble* using a genetic algorithm to find the optimal weights. We use this ensemble to produce the final forecasts. In total, we tested 33 thematic constituent models grouped into five different overarching themes:

- Constituent models using features from election-level data from the last held election in the VDEM election-level data set
- Constituent models using features from the VDEM country year data set
- Constituent models using features from the digital societies project
- Constituent models using features from the World Bank's World Development indicators
- Constituent models using a combination of features from the above mentioned data sources

A list of all constituent models and which indicators each model contains can be found in Appendix B2, which also details which theme each constituent model corresponds to.

For each model, the features are taken as the last value in the last available calendar year. I.e. for forecasts one calendar year into the future, the values are taken from December of the previous year and for forecasts two calendar years into the future, the values are taken from December two calendar years back in time. Most features are only updated annually, and in these cases the features are simply lagged one or two years respectively, but for features such as election-related variables the values are taken to be the last observed value in the last year with available data. Missing data are replaced by filling the last observed value forward. This means that the EVI forecasting system can be considered a medium-term forecasting tool which forecasts the structural risk of electoral violence, but which does not take into account election dynamics in the individual election (e.g. the candidates records on encouragement of violence, or riots and retaliation escalation loops).

We use all national level elections to either the presidency or the lower house (or combinations of elections which included at least one of these) in the period 1989-2023 (last available data) and coded by VDEM as the training data for the true forecasts, which we then produce for 2024 and 2025, i.e. two calendar years into the future. 2024 can thus be seen as a side-casted year since the model does not have access to the data from 2024 but we can evaluate how well the model did on the elections which have already been held. In the evaluation of the models (more below) we split the training data into different training and test periods to produce out of sample forecasts.

The process for producing the forecasts are detailed in Figure 1 below:

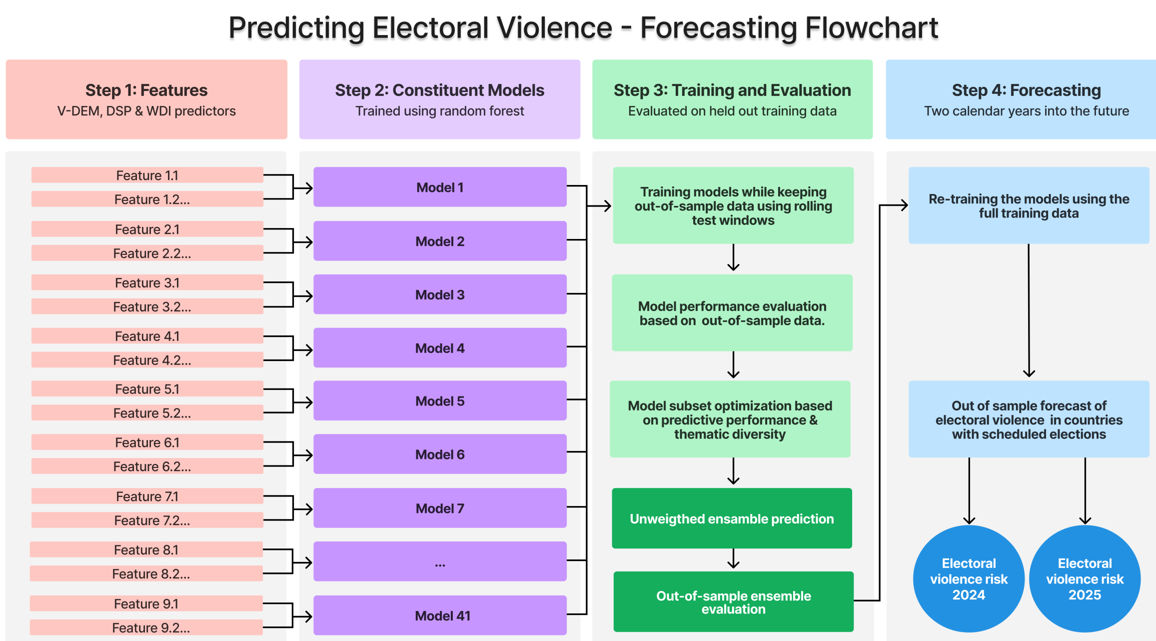

Figure 1. Forecasting flowchart

3.1 Forecasting models

To make the forecasts for electoral violence, we train the constitutent models using a random forest classifier(Breiman, 2001). Both of these models are developed to maximize the predictive performance of the model, rather than for optimizing the inferences which can be drawn from the models. This type of model can handle effects which are non-linear and effects which differ depending on other characteristics in the model (interactions). Models such as these have been shown to work well for predicting similar outcomes, for instance within the ViEWS project (Hegre et al., 2021). In the previous iteration of the EVI, we also used an extreme gradient boosted classifier (xgboost). However, due to similar predictive performance and other complexities in the xgboost model, such as the tuning of hyper parameters, we decided to proceed with only the random forest classifier for the EVI tool. The random forest classifier is less prone to overfitting than the xgboost model and is also less computationally demanding. More details on this change can be found under section on methdological changes for the EVI.

As the project uses a multitude of constituent models, these need to be weighted together in order to produce a final prediction. We do this by using a genetically weighted ensemble which optimizes the *Brier* score of the ensemble in the rolling test window. In total, the the genetic algorithm selects 12 models which are included in either the one- or two years ahead forecasts, of which 7 were selected for the one-year ahead forecast and 6 for the two-years ahead forecast. The models included in the ensemble, including their relative weights, are shown in table 1 in Appendix B1. The genetically weighted ensembles rely heavily on the irregularities and/or characteristics of the last election, including the reported level of electoral violence for the last election, with 54% and 61% of the final ensemble weights coming from these models in the one- and two-years ahead forecasts respectively. The remaining weights are distributed among models that include a range of different structural features, such as VDEM mid- and low-level indices, WDI structural indicators, and models which contain features from the digital society project (DSP). The use of a genetically weighted ensemble new in this year's iteration of the EVI as we previously used an unweighted ensemble of the best 9 models among our constituent models. The differences between the two methods are discussed in the section on methodological changes for the EVI.

3.2 Evaluation of models

The predictive performance of the constituent models and the ensembles were evaluated using a range of standard evaluation statistics. More specifically, the accuracy, brier, area under the precision recall curve (AUPR) and area under the receiver operator characteristic curve (AUROC) scores were computed. For the AUPR and AUROC scores, the scores are the unweighted averages (i.e. macro) of the three one-against-all AUPR/AUROC scores. Accuracy measures the proportion of cases which are correctly classified on our three-level ordinal scale when the most likely level of electoral violence is taken as the prediction. The brier score, on the other hand, is a measure of the squared error, in terms of (decimal) probability for the model. AUPR measures the performance of the model when trading off the precision, i.e. the proportion of correctly classified cases among predicted positive cases, with recall, i.e. the proportion of all positive cases correctly predicted. AUROC is similar to AUPR but measures the performance of the model when trading off the recall with the false positive rate, i.e. the proportion of predicted positive cases which are in fact negative. Accuracy, AUPR and AUROC all theoretically range from 0 to 1 where 0 is the worst performing model and 1 is the best performing model. The Brier score also range from 0 to 1, but for this score a lower value indicate better performance (Zhou et al., 2021).

Crucial when evaluating machine learning models such as the random forest models is that the evaluation happens on data which the model has not seen, since the models tend to overfit (i.e. perform artificially well) on data which are within the sample. To ensure that the constituent models were evaluated on out-of-sample data, we evaluated the performance using a rolling test window. Here, the period 2011-2023 was treated as the test period. For each year, the models were trained using all year prior, and predictions made one year into the future. I.e. for 2011 the period 1989-2010 was used as the training data, while for 2023 the period 1989-2022 was used as training data.

3.3 Evaluation results

The results of the evaluation of the individual constituent models as well as the genetically weighted algorithm can be seen in Appendix C Tables 7 and 8 for one and two year ahead predictions respectively. The lists in Appendix C1 are sorted on the *brier-score* metric, where a lower value indicates a better performance. The results show that the best performing models have accuracies between 77-83%, brier scores of 0.12-0.14. and AUPR and AUROC scores of 0.82-0.85 and 0.92-0.94 for both, the one and two year ahead forecasts.

Worth noting here is that the best performing models all include variables relating to the characteristics of the last held election, including the level of violence of that election. This is in line with results from other conflict forecasting efforts, such as the ViEWS project, where conflict history are usually the best predictors of future violence. The genetically weighted ensemble using has the best performance on the brier core, AUPR, and AUROC metrics and among the highest performance on the accuracy metrics.

In Figure 2 below are the performance metrics for the one and two year ahead forecasts in the rolling test window (rw1 and rw2) specification for the genetically weighted ensemble as well as the full digital society project model.

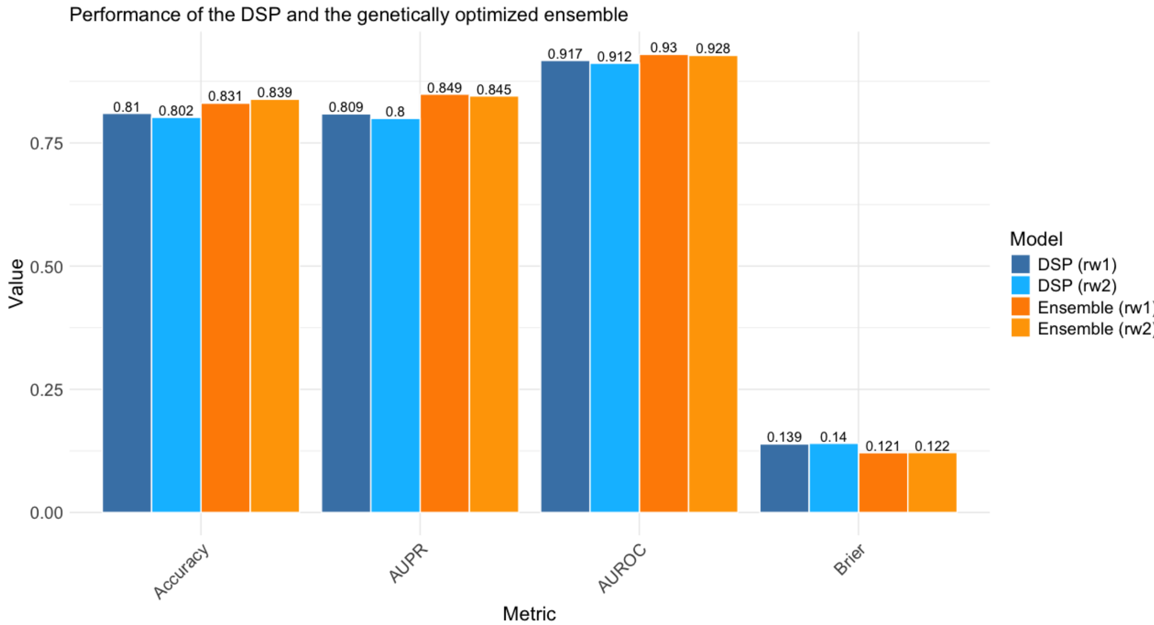

Figure 2. Performance comparison of random forest models for the one and two year ahead forecasts in the rolling test window

4 Methodological changes in the EVI

There are three main methodological changes in the 2024 iteration of the EVI compared to the initial 2022 version of the EVI. These are:

1. A slight tweaking of the definitions used to categorize electoral violence into the three categories. 2. A switch from using both a Random Forest classifier and an xgboost classifier 3. A changing in the final ensemble of mdoels used for the true forecasts from an unweighted average of the top nine models to a genetic algorithm which weights the models based on their performance.

4.1 Tweaking the definitions of electoral violence

In the 2024 iteration of the EVI, the threshold for elections considered to be peaceful has been lowered slightly such that when re-coding the two indicators on the five-point scale to the three point scale we use 3 as the lower limit for 'no electoral violence' rather than 3.5. In practice, this amounts to some elections which were previosly considered to have a limited amount of electoral violence being re-classified as peaceful. Among the 1,736 elections in the period 1989-2023 used for our training data, 197 are re-classified from 'limited' to 'no' electoral violence.

Motivation: We made this change as it better separates countries with peaceful elections from those with more visible electoral violence by letting the more ambiguous cases be classified as 'no electoral violence'. Our belief is that this classification better reflects the public conception of what electoral violence is and how it is expressed.

Consequences: Tweaking the definitions of electoral violence has, in general, only small effects on the performance of the models. However, as we re-classify some ambiguous elections as peaceful, we are also increasing the difficulty of the prediction task. This increased difficulty is noticable in the fact that the evaluation metrics are slightly worse for the 2024 iteration of the EVI compared to the 2022 version of the EVI. Another consequence of the tweaking of the definitions is that the risk index is not directly comparable between the 2022 and 2024 versions of the EVI. However, the differences are small and the rankings are generally similar, although the risk index is expected to generally be slightly lower in the 2024 version of the EVI.

4.2 Switching to using only Random Forest classifiers

In the 2022 iteration of the EVI, we used both a Random Forest classifier and an xgboost classifier to make the final predictions. In the 2024 iteration of the EVI, we have switched to only using the Random Forest classifier.

Motivation: We compared the performance of the two classifiers for the 2024 iteration of the EVI and found that the performance was similar for the two classifiers. As the Random Forest classifier is easier to interpret and has a lower risk of overfitting, we decided to only use the Random Forest classifier for the 2024 iteration of the EVI. The Random Forest classifier is also more robust to hyperparameter tuning, which makes it easier to use in practice.

Consequences: The switch to only using the Random Forest classifier has not had a major impact on the performance of the EVI. The performance of the Random Forest classifier is similar to the performance of the xgboost classifier, and the final ensemble of models used for the true forecasts has a similar performance to the ensemble used in the 2022 iteration of the EVI. We believe that the switch to only using the Random Forest classifier has made the EVI easier to use and interpret.

4.3 Changing the final ensemble of models

In the 2022 iteration of the EVI, we used an ensemble of 9 constituent models to make the final predictions. In the 2024 iteration of the EVI, we have switched to using an ensemble of 5 constituent models.

Motivation: The 9 constituent mdoel ensemble used in the 2022 iteration of the EVI was found to be somewhat redundant, as the models were highly correlated and did not add much to the final predictions. By using a genetic algorithm to create a weighted ensemble across all models we are able to make the ensemble more robust as it is less dependent on the performance of individual models. By using a genetic algorithm to create the ensemble, we are also able to better separate predictions from highly correlated models and make the ensemble more robust to overfitting. The genetic algorithm also allows us to use a smaller number of models in the ensemble, which makes the EVI easier to use and interpret.

Consequences: The switch to using a genetically weighted ensemble of constituent models has not had a major impact on the performance of the EVI. We compared the performance of the 9 constituent model ensemble and genetically weighted ensemble for the 2024 iteration of the EVI and found that the performance was similar for the two ensembles, but with a slight advantage for the genetically weighted algorithm. We believe that the switch to using a genetically weighted ensemble is a more principled choice for the EVI and makes the EVI more robust.

References

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45, pp. 5–32.

Coppedge, Michael et al. (2021a). *V-Dem Codebook v11*.

Coppedge, Michael et al. (2021b). *V-Dem Country-Year/Country-Date Dataset v11*.

Fjelde, Hanne et al. (2021). *The Deadly Electoral Conflict Dataset (DECO) Codebook*.

Hegre, Håvard et al. (2019). “ViEWS: A political violence early-warning system”. In: *Journal of peace research* 56.2, pp. 155–174.

Hegre, Håvard et al. (2021). “ViEWS2020: revising and evaluating the ViEWS political violence early-warning system”. In: *Journal of peace research* 58.3, pp. 599–611.

Mechkova, Valeriya et al. (2022). *Digital Society Project Dataset v4*.

Zhou, Jianlong et al. (2021). “Evaluating the quality of machine learning explanations: A survey on methods and metrics”. In: *Electronics* 10.5, p. 593.

Appendix A: The outcome of electoral violence

Definition of v2elintim:

Question: In this national election, were opposition candidates/parties/campaign workers subjected to repression, intimidation, violence, or harassment by the government, the ruling party, or their agents?

Clarification: Other types of clearly distinguishable civil violence, even if politically motivated, during the election period should not be factored in when scoring this indicator (it is dealt with separately).

Responses:

0: Yes. The repression and intimidation by the government or its agents was so strong that the entire period was quiet.

1: Yes, frequent: There was systematic, frequent and violent harassment and intimidation of the opposition by the government or its agents during the election period.

2: Yes, some. There was periodic, not systematic, but possibly centrally coordinated — harassment and intimidation of the opposition by the government or its agents.

3: Restrained. There were sporadic instances of violent harassment and intimidation by the government or its agents, in at least one part of the country, and directed at only one or two local branches of opposition groups.

4: None. There was no harassment or intimidation of opposition by the government or its agents, during the election campaign period and polling day.

Definition of v2elpeace:

Question: In this national election, was the campaign period, election day, and post-election process free from other types (not by the government, the ruling party, or their agents) of violence related to the conduct of the election and the campaigns (but not conducted by the government and its agents)?

Responses:

0: No. There was widespread violence between civilians occurring throughout the election period, or in an intense period of more than a week and in large swaths of the country. It resulted in a large number of deaths or displaced refugees.

1: Not really. There were significant levels of violence but not throughout the election period or beyond limited parts of the country. A few people may have died as a result, and some people may have been forced to move temporarily.

2: Somewhat. There were some outbursts of limited violence for a day or two, and only in a small part of the country. The number of injured and otherwise affected was relatively small.

3: Almost. There were only a few instances of isolated violent acts, involving only a few people; no one died and very few were injured.

4: Peaceful. No election-related violence between civilians occurred.

Definition of electoral violence target:

Recoding of v2elintim and v2elpeace into three level ordinal variables with values [0-1.5] corresponding to 'Severe electoral violence/intimidation', values (1.5-3) corresponding to 'limited electoral violence/intimidation', and value [3-4] corresponding to 'no electoral violence/intimidation' perpetrated by government actors (v2elintim) and non-government actors (v2elpeace) respectively. The target 'electoral violence' is then taken as the max of the two three level ordinal variables measuring electoral violence/intimidation.

Appendix B1: Constituent models in the final ensemble

The final ensemble for the predictions consists of an genetically weighted ensemble. The ensembles rely heavily on the irregularities and/or characteristics of the last election, including the reported level of electoral violence for the last election, with 54% and 61% of the final ensemble weights coming from these models in the one- and two-years ahead forecasts respectively. The remaining weights are distributed among models that include a range of different structural features, such as VDEM mid- and low-level indices, WDI structural indicators, and models which contain features from the digital society project (DSP). The exact constituent models and their weights in the one- and two-years ahead forecasts are shown in table 1 below.

Table 1. Weights of the constituent models in the one and two year ahead ensembles, respectively. Ordered by the total weight in either ensembles.

Constituent model	w_{1yr}	w_{2yr}
Election Irregularities last election (short)	0.00	0.47
Election Irregularities last election (long)	0.25	0.14
Election characteristics last election (long)	0.29	0.00
VDEM mid level indicies and WDI structural	0.00	0.13
VDEM civil liberties indicies	0.00	0.11
VDEM mid level indicies, WDI structural, and DSP infrastructure	0.10	0.00
Full VDEM, WDI, and DSP model	0.10	0.00
Full VDEM model	0.09	0.00
VDEM mid level indicies (alternative)	0.09	0.00
DSP social media climate and usage	0.00	0.08
Full DSP model	0.08	0.00
VDEM accountability indicies	0.00	0.07

Appendix B2: Constituent models description

Below are tables describing the 33 thematic constituent models evaluated for the final ensemble, divided into five categories: 1) constituent models focusing on the characteristics of the previous election; 2) constituent models using the Digital Society Project indicators for digital infrastructure and vulnerability; 3) constituent models focusing on VDEM yearly indicators; 4) constituent models using the World Development Indicators; and 5) combination models which mix features across the four data sources.

Table 2. Election history (VDEM-CD) constituent models

Model name	Description of features	Included features
History of electoral violence (history only)	Features tracking streaks of peaceful, severely violent, and low-violence elections, and # elections since the last constitutional change	cons_elect, peaceful_streak, violent_streak, lowviolent_streak
History of electoral violence (full)	Features from history of electoral violence model, and reported levels of electoral violence in the last election	"History of electoral violence (history only)" plus v2elintim_osp, v2elpeace_osp
Election Irregularities last election (short)	Irregularity-related features from the last election	v2elembaut, v2elembcap, v2elmulpar, v2elvotbuy, v2elirreg, v2elintim, v2elpeace, v2elboycot, v2elfrfair, v2elmonden, v2elmonref
Election Irregularities last election (long)	Irregularity-related features from the last election, including violence streaks	"Election Irregularities last election (short)" plus cons_elect, peaceful_streak, violent_streak, lowviolent_streak
Election Characteristics last election (structural)	Structural features from the last election	v2asuffrage, v2elcomvot, v2elgvsuflvl, v2eldonate, v2elpubfin, v2elembaut, v2elembcap, v2elmulpar, v2elrgstry, v2elvotbuy, v2elfrcamp, v2elpdcamp, v2elpaidig, v2eldommon, v2elintmon, v2elvaptrn
Election Characteristics last election	Characteristics of the last election features from the V-Dem country-date dataset	"Election Characteristics last election (structural)" plus v2elirreg, v2elintim, v2elpeace, v2elboycot, v2elfrfair, v2elmonden, v2elmonref, v2elaccept, v2elasmoff, cons_elect
Election Characteristics last election (full)	Characteristics of the last election features from the V-Dem country-date dataset, including violence streaks	"Election Characteristics last election" plus peaceful_streak, violent_streak, lowviolent_streak

Table 3. Digital Society Project constituent models

Model name	Description of features	Included features
DSP Monitoring	DSP features relating to government monitoring, surveillance, and repression online	v2smregcap, v2smgovfilprc, v2smgovsmmon, v2smgovsmcenprc, v2smarrest
DSP Disinformation and social media usage	DSP features relating to disinformation online and social media usage	v2smgovdom, v2smpardom, v2smfordom, v2smorgelitact, v2smcamp
DSP Social media climate	Social media climate from DSP, including dissemination of disinformation, online polarization and hate speech, and traditional use of social media by elites/political candidates	"DSP Disinformation and social media usage" plus v2smonper, v2smmefra, v2smpolsoc, v2smpolhate
DSP Social Media Climate, security	DSP features relating to social media climate, security, and usage	"DSP Social media climate" plus v2smgovcapsec, v2smpolcap
DSP Infra	Digital infrastructure features, including media features from V-Dem-CY, embassy capacity from V-Dem-CY, cyber security + monitoring and surveillance of social media from DSP, and internet use from WDI	"DSP Monitoring" plus v2smonex, v2elfrcamp, v2mecrit, v2merange, v2elembaut, it.net.user.zs
DSP Disinformation and social climate and usage	DSP features relating to disinformation online and social media climate and usage	"DSP Social Media Climate, security"
DSP full model	All interval scale features from DSP	"DSP Disinformation and social climate and usage" plus v2smgovab, v2smparab, v2smforads, v2smgovfilcap, v2smgovshutcap, v2smgovshut, v2smgovsm, v2smgovsmalt, v2smgovcapsec, v2smregcon, v2smprivex, v2smprivcon, v2smregapp, v2smlawpr, v2smdefabu, v2smonex, v2smorgviol, v2smorgavgact

Table 4. VDEM Country-Year constituent models

Model name	Description of features	Included features
VDEM Political Exclusion Indicies	VDEM-CY features on exclusion of groups	v2xpe_exlecon, v2xpe_exlgender, v2xpe_exlgeo, v2xpe_exlpol, v2xpe_exlsocgr
VDEM Neopatrimonialism	VDEM-CY neopatrimonialism features	v2x_neopat, v2xnp_client, v2xnp_pres, v2xnp_regcorr
VDEM Civil Liberties Indicies	VDEM-CY features on civil liberties	v2x_clphy, v2x_clpol, v2x_clpriv, v2x_civlib
VDEM Accountability Indicies	VDEM-CY features on accountability	v2x_accountability, v2x_veracc, v2x_diagacc, v2x_horacc
VDEM Gender	VDEM-CY gender features	v2x_gencl, v2x_gencs, v2x_genpp, v2x_gender
VDEM High level indicies	VDEM-CY high-level indices	v2x_polyarchy, v2x_libdem, v2x_partipdem, v2x_delibdem, v2x_egaldem
VDEM mid level indicies (alternative)	VDEM-CY mid-level indices	"VDEM Accountability Indicies" plus v2x_neopat, v2x_civlib, v2x_gender, v2x_corr, v2x_rule, v2xcs_ccsi, v2xps_party, v2x_divparctrl, v2x_feduni
VDEM mid level indicies	VDEM-CY and CD mid-level component indices	"VDEM High level indicies" plus v2x_api, v2x_mpi, v2x_freexp_altinf, v2x_frassoc_thick, v2x_suffr, v2xel_frefair, v2x_elecoff, v2xcl_rol, v2x_jucon, v2xlg_legcon, v2x_cspart, v2xdd_dd, v2xel_locelec, v2xel_regelec, v2xdl_delib, v2xeg_eqprotec, v2xeg_eqaccess
VDEM full model	All v2x indices from VDEM-CY	"VDEM mid level indicies" plus v2x_ex_confidence, v2x_ex_direlect, v2x_ex_hereditary, v2x_ex_military, v2x_ex_party, v2xnp_client, v2xnp_pres, v2xnp_regcorr, v2xdd_cic, v2xdd_i_ci, v2xdd_i_rf, v2xdd_toc, v2xdd_i_pl, v2xdd_i_or, v2xcs_ccsi, v2x_EDcomp_thick, v2xcl_disc, v2xcl_dmove, v2xcl_slave, v2xex_elecleg, v2xme_altinf, v2xps_party, v2x_divparctrl, v2x_feduni, v2xca_academ

Table 5. World Development Indicators Constituent model

Model name	Description of features	Included features
WDI Education	WDI education factors, including enrollment and expenditure	se.enr.prim.fm.zs, se.enr.prsc.fm.zs, se.prm.nenr, se.xpd.totl.gb.zs, se.xpd.totl.gd.zs
WDI Resources	WDI factors on resources and GDP	ny.gdp.pcap.kd, ny.gdp.pcap.kd.zg, dt.oda.odat.pc.zs, ny.gdp.petr.rt.zs, ny.gdp.totl.rt.zs
WDI Structural	WDI structural factors, including population, age composition, IMR, life expectancy, and GDP	sp.pop.totl, ny.gdp.pcap.kd, ny.gdp.pcap.kd.zg, sp.dyn.imrt.in, sp.dyn.le00.in, sp.pop.0014.fe.zs, sp.pop.grow, sp.pop.65up.fe.zs
WDI full	WDI full model	"WDI Education", "WDI Resources", and "WDI Structural" plus ms.mil.xpnd.zs, ms.mil.xpnd.gd.zs, nv.agr.totl.kn, sp.dyn.le00.in, sh.sta.maln.zs, sh.sta.stnt.zs, sl.tlf.totl.fe.zs, sm.pop.totl.zs, sh.dyn.mort.fe, sp.pop.1564.fe.zs, sp.urb.totl.in.zs, sl.uem.neet.zs, it.net.user.zs

Table 6. Combination Constituent models

Model name	Description of features	Included features
VDEM High level indicies and WDI structural	Combination of features from VDEM High level indices and WDI structural	"VDEM High level indicies" and "WDI Structural"
VDEM Mid level indicies and WDI structural	Combination of features from VDEM mid level indices and WDI structural	"VDEM mid level indicies" and "WDI Structural"
Election Irregularities (last election), VDEM civil liberties, and WDI structural	Combination of features from VDEM Civil Liberties, election irregularities (last election), and WDI structural	"VDEM Civil Liberties Indicies", "Election Irregularities last election (short)", and "WDI Structural"
Election Irregularities (last election), VDEM exclusion, and WDI structural	Combination of features from VDEM Political Exclusion, election irregularities (last election), and WDI structural	"VDEM Political Exclusion Indicies", "Election Irregularities last election (short)", and "WDI Structural"
VDEM Mid level indicies, WDI structural, and DSP infrastructure	Combination of features from VDEM mid level indices, WDI structural, and DSP infrastructure	"VDEM Mid level indicies and WDI structural" and "DSP Infra"
Full model	Combination of all features above	"VDEM Mid level indicies, WDI structural, and DSP infrastructure", "Election Irregularities (last election), VDEM civil liberties, and WDI structural", and "Election Irregularities (last election), VDEM exclusion, and WDI structural"

Appendix C: Model Performance

Rank	Model	Accuracy	Brier	AUROC	AUPR
1	Genetically optimized ensemble	0.831	0.121	0.930	0.849
2	Election Irregularities (last election), VDEM exclusion and WDI structural	0.843	0.123	0.905	0.807
3	Election Characteristics last election (full)	0.834	0.124	0.921	0.839
4	Election Irregularities (last election), VDEM civil liberties, and WDI structural	0.850	0.125	0.907	0.812
5	Election Irregularities last election (long)	0.831	0.126	0.920	0.833
6	Election Irregularities last election (short)	0.833	0.126	0.920	0.835
7	Election Characteristics last election (full)	0.836	0.126	0.919	0.837
8	VDEM Mid level indicies, WDI structural, and DSP infrastructure	0.833	0.129	0.914	0.815
9	History of electoral violence (full)	0.829	0.131	0.914	0.821
10	VDEM Mid level indicies and WDI structural	0.824	0.131	0.909	0.805
11	VDEM full model	0.829	0.132	0.919	0.832
12	VDEM mid level indicies	0.817	0.138	0.916	0.841
13	VDEM High level indivies and WDI structural	0.824	0.139	0.903	0.807
14	DSP full model	0.810	0.139	0.917	0.809
15	VDEM mid level indicies	0.805	0.145	0.902	0.802
16	DSP Social Media Climate, security	0.801	0.150	0.905	0.804
17	Election Characteristics last election, structural	0.798	0.151	0.895	0.806
18	DSP Infra	0.788	0.151	0.889	0.780
19	DSP Disinformation and social media usage	0.801	0.156	0.892	0.786
20	DSP Disinformation and social climate and usage	0.797	0.156	0.892	0.787
21	VDEM High level indicies	0.781	0.157	0.891	0.788
22	DSP Social media climate	0.801	0.157	0.891	0.784
23	VDEM Political Exclusion Indicies	0.781	0.158	0.893	0.797
24	History of electoral violence (history only)	0.779	0.160	0.883	0.763
25	Full model (all features)	0.781	0.166	0.885	0.781
26	VDEM Neopatrimonialism	0.747	0.171	0.864	0.750
27	WDI Structural	0.743	0.174	0.853	0.701
28	DSP Monitoring	0.761	0.176	0.875	0.746
29	VDEM Accountability Indicies	0.751	0.176	0.862	0.760
30	VDEM Civil Liberties Indicies	0.735	0.186	0.850	0.734
31	WDI full model	0.758	0.204	0.807	0.677
32	VDEM Gender	0.697	0.215	0.820	0.682
33	WDI Education	0.673	0.217	0.758	0.617
34	WDI Resources	0.650	0.233	0.788	0.622

Table 7. Performance of models in the one year ahead prediction task

Rank	Model	Accuracy	Brier	AUROC	AUPR
1	Genetically optimized ensemble	0.839	0.122	0.928	0.845
2	Election Irregularities last election (short)	0.843	0.124	0.923	0.841
3	Election Irregularities (last election), VDEM exclusion, and WDI structural	0.841	0.124	0.906	0.808
4	Election Irregularities last election (long)	0.839	0.125	0.918	0.836
5	Election Characteristics last election (full)	0.836	0.126	0.917	0.828
6	Election Irregularities (last election), VDEM civil liberties, and WDI structural	0.841	0.126	0.906	0.812
7	Election Characteristics last election (full)	0.844	0.126	0.919	0.831
8	VDEM Mid level indicies, WDI structural, and DSP infrastructure	0.819	0.132	0.905	0.813
9	History of electoral violence (full)	0.811	0.133	0.912	0.821
10	VDEM Mid level indicies and WDI structural	0.826	0.134	0.903	0.804
11	VDEM full model	0.823	0.136	0.915	0.816
12	VDEM High level indivies and WDI structural	0.798	0.140	0.900	0.812
13	VDEM mid level indicies	0.807	0.140	0.912	0.830
14	DSP full model	0.802	0.140	0.912	0.800
15	DSP Social Media Climate, security	0.789	0.148	0.903	0.802
16	DSP Infra	0.805	0.148	0.890	0.785
17	Election Characteristics last election, structural	0.800	0.149	0.895	0.795
18	VDEM mid level indicies	0.799	0.149	0.897	0.794
19	DSP Disinformation and social media usage	0.807	0.151	0.897	0.800
20	DSP Disinformation and social climate and usage	0.799	0.152	0.896	0.795
21	DSP Social media climate	0.801	0.153	0.893	0.791
22	VDEM Political Exclusion Indicies	0.771	0.160	0.886	0.773
23	VDEM High level indicies	0.765	0.160	0.891	0.791
24	Full model (all features)	0.767	0.166	0.886	0.791
25	History of electoral violence (history only)	0.759	0.167	0.874	0.751
26	WDI Structural	0.760	0.173	0.856	0.708
27	VDEM Accountability Indicies	0.757	0.177	0.859	0.746
28	VDEM Neopatrimonialism	0.733	0.180	0.855	0.729
29	DSP Monitoring	0.741	0.181	0.865	0.727
30	VDEM Civil Liberties Indicies	0.729	0.192	0.838	0.717
31	WDI full model	0.735	0.203	0.814	0.676
32	VDEM Gender	0.701	0.210	0.819	0.689
33	WDI Education	0.661	0.220	0.749	0.606
34	WDI Resources	0.653	0.238	0.779	0.614

Table 8. Performance of models in the two year ahead prediction task