Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04967-w OPEN

Predicting police and military violence: evidence from Colombia and Mexico using machine learning models

Juan David Gelvez[1]✉

Armed forces violence has pervasive effects on public trust and population well-being. Such misconduct is not random, making its prevention both crucial and challenging due to the difficulty of measuring and detecting these phenomena beforehand. Recent advances in artificial intelligence offer new tools for this task. This article proposes the use of machine learning models to predict armed forces violence at the municipality level. Focusing on Colombia and Mexico—two countries with a significant number of human rights abuses by armed forces—the analysis draws on comprehensive subnational datasets. In Colombia, the study examines 1255 extrajudicial killing cases in which innocent civilians were misrepresented as guerrillas by the military. In Mexico, it considers 12,437 allegations of severe human rights abuses during militarized policing operations. Separate machine learning models are trained using four canonical algorithms—Lasso, Random Forests, Extreme Gradient Boosting, and Neural Networks—and their predictions are combined through a Super Learner ensemble. Results show high accuracy, specificity, and sensitivity in predicting police and military violence. In addition, feature-importance analysis highlights the most influential variables in the models' predictions. These findings carry significant policy implications for contemporary law-and-order institutions, particularly in Latin America, where over a quarter of the world's homicides occur, less than half the population expresses confidence in the police, and more than 9000 police killings are reported in a single year.

[1] University of Maryland, College Park, College Park, MD, USA. ✉email: jgelvez@umd.edu

Introduction

Military and Police misconduct poses significant risks to institutional stability and citizen welfare. From high-profile scandals of extrajudicial executions to the everyday abuses that go unreported, such misconduct not only undermines the armed forces' ability to protect the rule of law but also sows distrust, corruption, and fear within communities. In fact, countries with high levels of reported human rights violations often experience stalled social and economic development (Apergis and Cooray 2020), eroding the foundations of institutional trust (Curtice 2021; Sung et al. 2022).

Yet, the full scope of military and police misconduct remains largely invisible. These actions often occur in secrecy, and detection typically relies on victims or witnesses—if they exist—willing to come forward. As a result, traditional tools like perception surveys, where civilians, officers, and experts offer their opinions on misconduct, fall short in predicting when and where abuses will occur (See, for instance, Kutnjak Ivkovic 2005 and Woolfolk et al. 2021).

This paper aims to fill this gap by employing predictive models to anticipate misconduct within the armed forces[1] and identify the key features that drive such behavior. I build several machine learning models and a Super Learner ensemble to predict municipality-level misconduct by armed forces using cases from Colombia (2000–2010) and Mexico (2006–2016). In Colombia, I analyze the location of 1255 extrajudicial executions, where innocent civilians were killed and misrepresented as guerrillas by the Colombian military. The location of extrajudicial executions comes from the meticulous collecting efforts of a Colombian Jesuit NGO, based on direct reports from the ground, including from the clergy, and detailed analysis of various national and local news sources, and organized by Acemoglu et al. (2020). For Mexico, I focus on serious human rights abuse complaints during militarized policing deployments, which include arbitrary detention, extrajudicial killings, and torture, based on data from 12,437 documented cases of serious abuses. This information comes from an important effort by Flores-Macías and Zarkin (2024) after winning a series of appeals before Mexican public institutions.

To enhance the prediction of military misconduct, my research incorporates machine learning methods characterized by their innovative nature and capability to handle complex datasets. Central to this approach is the use of a stacking learning strategy, which integrates the output of various base models to formulate a consolidated and more precise final prediction. This technique leverages the strengths of multiple predictive models to reduce bias and variance, resulting in improved prediction accuracy; and it is highly used and well-regarded in the literature on prediction of misconduct and deviant behavior (see, for example, Berk 2017; Cubitt and Birch 2021; Cubitt et al. 2022; Jenasamanta and Mohapatra 2022; Bazzi et al. 2022; Gallego et al. 2022; Low et al. 2024). Specifically, I have developed separate models for each country of interest—employing a diverse array of canonical machine learning algorithms, including Lasso, Random Forests, Extreme Gradient Boosting, and Neural Networks. Each model brings a unique perspective, capturing different aspects and dynamics of potential misconduct.

Using a feature-importance analysis, I organize the predictor variables into four categories: (1) Socioeconomic and Demographic Factors, including population size, unemployment, and quality of education; (2) Political and Institutional Context, such as judicial efficiency, political attitudes, and political alignment; (3) Military and Security Factors, covering variables like violence, military presence, and security operations; and (4) Geographic and Environmental Conditions, which account for factors like accessibility, and rurality. This setup also allows to identify the categories with the greatest predictive power to anticipate where misconduct might occur.

This paper presents compelling evidence that military and police misconduct can be reliably anticipated at the municipal level. By leveraging machine learning methods—particularly the SuperLearner algorithm—even the most conservative estimates achieve over 92% predictive accuracy, with a balanced sensitivity of 87% and specificity of 84%. This approach allows for the identification of areas and contexts at high risk for human rights abuses committed by armed forces. While this analysis is predictive and not intended to establish causal relationships, the feature importance analysis reveals some trends: in Colombia, geographic and environmental factors are the most influential, whereas in Mexico, socioeconomic and demographic variables play a more significant role in the prediction process.

Given these insights, this research holds significant implications. This study is one of the first to integrate machine learning models to predict armed forces misconduct, particularly at the municipal level in two distinct national security contexts: Colombia and Mexico. By leveraging comprehensive sub-national datasets and advanced predictive techniques, it also bridges a critical gap in the literature on preventing human rights abuses by security forces.[2] Likewise, as highlighted in the conclusion, the results demonstrate the efficacy of machine learning techniques in reducing the human and institutional costs associated with such abuses, thereby enhancing public trust and accountability. These findings can inform policymakers in developing strategies to reduce misconduct and promote institutional reforms that strengthen governance and protect human rights in the region.

Armed forces misconduct

Police and military misconduct refer to actions or behaviors by armed forces personnel that violate legal, ethical, or professional standards, undermining their fundamental role in maintaining national security and enforcing the rule of law (Kappeler et al. 1998). Misconduct can range from overt abuses, such as extrajudicial killings, to systemic failures, including complicity in organized crime. These actions compromise their capacity to fulfill essential duties (Blair and Weintraub 2023), often resulting in significant harm to both institutional stability (Greitens 2016) and citizen welfare (Lawrence 2017). Misconduct in the police has far-reaching effects, eroding public trust in institutions and fostering an environment of impunity (González 2020; Gelvez et al. 2022; Salazar-Tobar and Rengifo 2023), particularly in regions where law enforcement is weak or corrupt (Sung et al. 2022; Gingerich and Oliveros 2018).

The consequences of military misconduct extend beyond immediate victims. As resources are misallocated and human rights are violated, countries often experience weakened institutional governance (Curtice 2021), slower economic growth and increase poverty (Apergis and Cooray 2020), and diminished social cohesion (Blair et al. 2022). The perpetuation of misconduct creates a feedback loop of corruption, discouraging victims from coming forward and inhibiting accountability mechanisms (Gingerich and Oliveros 2018).

Detecting military misconduct presents substantial challenges due to its clandestine nature, often occurring in environments with limited oversight, which complicates observation and documentation (Rowe 2008). Nevertheless, existing literature has identified key predictors of armed forces misconduct. To capture these events, I categorize explanatory factors into four groups that have been used to analyze police and military misbehavior. The next subsection examines these categories in detail from a theoretical perspective.

Predictors of police and military misconduct. Armed forces misconduct is rarely a random occurrence; rather, it arises from a complex interaction of structural, political, security, and geographic factors. The literature across political science, criminology, sociology, and other behavioral disciplines, identifies key categories that influence misconduct, which can be grouped into four broad predictor groups: i) socioeconomic and demographic factors; ii) political and institutional context; iii) military and security conditions; and iv) geographic and environmental characteristics. Each category captures unique influences that increase or mitigate the likelihood of misconduct, offering valuable insights for predictive modeling.

First, socioeconomic and demographic factors play a crucial role in shaping the environment in which armed forces operate. Higher levels of poverty, unemployment, and inequality make populations more vulnerable to violence and military abuses (Pridemore 2011; Evans and Kelikume 2019; Gelvez and Johnson 2023; Franc and Pavlovic 2023). In line with social disorganization theories, structural disadvantage —characterized by poverty, low education levels, and economic deprivation— creates conditions conducive to misconduct by weakening informal social controls and fostering police-citizen conflict (Kane 2002). For example, when the military is deployed to restore order in economically disadvantaged areas, the likelihood of misconduct increases, as soldiers may face stressful conditions (Caforio 2014) or be tasked with policing roles for which they are not adequately trained (Blair and Weintraub 2023). Similarly, economic and social inequality can foster resentment between marginalized populations and state institutions (Blair et al. 2022), creating fertile ground for misconduct. In regions where education levels are low, both the general population may lack awareness of human rights and legal frameworks, further increasing the probability of misconduct. In contrast, higher levels of education within the population often correlate with stronger civilian oversight and democratic norms, which can help curb abuses (Stone and Ward 2000).

Second, the political and institutional context is central to understanding armed forces misconduct. Strong, functioning institutions typically enforce accountability and deter misconduct, while weak or corrupt institutions create opportunities for abuses to flourish. Research shows, for example, that supervisory oversight and organizational discipline mechanisms, such as addressing civilian complaints and holding officers accountable, can significantly reduce future misconduct by improving officer behavior and deterring deviant subcultures (Lee et al. 2013; Rozema and Schanzenbach 2023). Judicial efficiency, therefore, plays a significant role in deterring military personnel from engaging in misconduct. Hu and Conrad (2020) shows that establishing judicial bodies for citizens to report allegations of police abuse provides "fire-alarm" oversight, enabling the monitoring of police officers for power abuses and reducing human rights violations by the police.

In addition, political attitudes shaped how security forces behaved. In regions where there is political alignment between national and subnational governments, national armed forces often operate with greater autonomy, facing fewer institutional checks and, therefore, a higher likelihood of impunity. As Flores-Macías and Zarkin (2021) argue, the constabularization of the military—where armed forces take on civilian policing roles—leads to an increased use of force and undermines efforts to reform civilian law enforcement. This militarization of public security creates a feedback loop where accountability weakens, and misconduct becomes more entrenched (González 2020). Similarly, Visconti (2019) shows how exposure to crime can shift public policy preferences toward more repressive crime-reduction measures, often at the expense of democratic norms. Building on this, Masullo and Morisi (2023) demonstrate that while citizens in crime-ridden contexts initially support the militarization of security forces, this support diminishes significantly when military operations result in civilian casualties, revealing a conditional and fragile basis for such preferences.

These dynamics align with public attitudes being further influenced by how protests are handled. When security forces respond to protests with violence, as shown by Nagel and Nivette (2023), public perceptions of law enforcement tend to shift negatively, with citizens seeing the police and military as politicized and detached from democratic norms. Such actions erode trust not only in law enforcement but in political institutions more broadly, further alienating the public and weakening democratic governance.

The third key category is military and security factors, which are directly related to the operational context of military deployments. Military presences in regions with weak oversight are more prone to abuses, as troops may operate without the necessary checks on their behavior, especially if their commanders have (lack of) incentives to control them (Bedi 2015; Acemoglu et al. 2020). Prolonged deployments, especially those involving policing roles, often lead to a breakdown in discipline, increasing the risk of misconduct (Blair and Weintraub 2023). Police involvement in counterinsurgency operations further heightens this risk (Gelvez et al. 2022). In these roles, the military is often tasked with controlling populations perceived as adversaries, which can lead to excessive force (Ortiz-Ayala 2021), arbitrary detentions, and extrajudicial killings. The intensity of conflict also plays a significant role. In regions experiencing high levels of violence or insurgency, military personnel are often under greater stress (Elbogen et al. 2014), and the rules of engagement may be relaxed, making misconduct more likely. In such environments, violence can become rationalized as a necessary tool for maintaining order, reflecting broader societal discourses on the legitimacy of force (Stroud 2020), which further complicates efforts to hold individuals accountable.

Geographic and environmental conditions significantly influence the likelihood of misconduct by affecting the ability of authorities to monitor military activities and enforce accountability measures. In remote conflict zones, fragmented governance and limited service provision hinder efforts to oversee military behavior and hold forces accountable for abuses (Kalyvas 2006). These areas often allow military forces to operate with a high degree of autonomy and face minimal scrutiny from higher authorities (Acemoglu et al. 2020). This autonomy is further exacerbated by the absence of media and civil society actors, which reduces opportunities for victims to report abuses and weakens accountability mechanisms (Campbell and Valera 2020). However, urban areas may also face heightened risks of human rights abuses. In densely populated regions, the proximity of civilians can intensify the potential for violations, as military operations intersect with daily civilian life (Pion-Berlin 2017).

Context and data

Extrajudicial executions in Colombia (2000–2010). Colombia has a long history of civil war and multiple non-state armed groups. The conflict involving the country's two largest guerrilla groups, the Revolutionary Armed Forces of Colombia (FARC) and the National Liberation Army (ELN), dominated the 2002 presidential election, which Álvaro Uribe won with his flagship policy, the Democratic Security Policy. Following years of mounting pressure to combat illegal groups under previous administrations, Uribe's approach involved significantly expanding the military and providing stronger incentives to

confront the guerrillas, bolstering his popularity by addressing citizens' concerns about the internal armed conflict (García-Sánchez and Rodríguez-Raga 2019). However, a significant repercussion of these heightened incentives was a surge in extrajudicial executions —instances where civilians were falsely labeled as guerrilla combatants and then killed by the army to receive rewards. Extrajudicial killings, popularized by the media as "Falsos Positivos", while not new to Colombia (Knoester 1998), increased dramatically following President Uribe's counter-insurgency strategy (Rodríguez Gómez 2020). This military practice was widespread throughout the country, not limited to isolated military units (Alston 2010; Aranguren Romero et al. 2021), and only began to decline after media reports revealed the extent of civilian deaths in 2008 (Acemoglu et al. 2020). Between 2002 and 2008, around 6000 innocent civilians were murdered and labeled as insurgents (Trust Commission 2024).

These extrajudicial killings exposed deep institutional weaknesses in Colombia's military command structure and judicial oversight. Research has shown that the root of these extrajudicial killings was tied to the military's incentives structure, where officers were rewarded for presenting high body counts as victories against insurgents. Colonels, in particular, who were in charge of brigades, were highly motivated to inflate these numbers to secure career promotions, while generals faced fewer career-related incentives to do so (Acemoglu et al. 2020). This institutional pressure fostered an environment where abuses flourished. In addition, municipalities with weaker judicial institutions were more prone to experiencing extrajudicial killings because military units could operate with minimal oversight. As Gordon (2017) argues, the overlap of high-powered incentives and the socio-economic inequalities made it easier for this misconduct to take place undetected. Aranguren Romero et al. (2021) further emphasize how these incidents disproportionately targeted marginalized populations, treating their lives as disposable under the guise of the war against insurgency.

To predict the sub-national location of military misconduct, I build a municipality-level panel dataset on the annual incidence of extrajudicial killings as a dummy-outcome variable. Though measuring misconduct is challenging, I use data made public by Acemoglu et al. (2020), which is fairly reliable. As explained by the authors, this data comes from the meticulous efforts of a Colombian Jesuit NGO, which collected direct reports from the ground and conducted detailed analyses of various national and local news sources. The data collected by Acemoglu et al. (2020) has been utilized in prior studies of violence (Albarracín et al. 2023) and even the prediction of conflict (Bazzi et al. 2022).

Furthermore, as predictor variables, I use annual municipal panel data, most of which is published by the Universidad de Los Andes in Colombia.[3] The dataset includes key variables relevant to understanding military misconduct, organized into the four predictor groups mentioned above. First, socioeconomic and demographic factors, such as annual measures of population size, unemployment rates, test scores in mathematics, language, and science, historical inequality measures, and the presence of Afro and Indigenous communities, capture the local context. Second, political and institutional context variables, including the strength of local judicial institutions and the extent of social mobilization, provide insights into local oversight and accountability mechanisms. Third, military and security factors, such as the presence of military brigades, levels of coca cultivation, and historical levels of violence, offer information on the operational environment in which the military acted. Finally, geographic and environmental conditions, including terrain characteristics like erosion, rurality, water availability, altitude, and access to infrastructure (e.g., distance to market), help assess how remoteness and physical characteristics of the area may have influenced military behavior. All variables in the dataset are measured annually at the municipality level. For more information about the data description, see the Supplementary Appendix.

Human rights violations in Mexico (2000–2016). In Mexico, the militarization of law enforcement became a central strategy in the government's fight against organized crime, particularly since 2006 when President Felipe Calderón declared an all-out war against drug cartels (Ley 2018). Although Mexico's Federal Police and Military had been involved in drug control since the 1960s, its role expanded dramatically during the war on drugs. This policy marked a turning point in the involvement of the armed forces in domestic policing operations, known as the process of constabularization, where the military took on roles traditionally held by civilian and police forces (Flores-Macías and Zarkin 2021, 2024). The Calderón administration's (2006–2012) strategy involved widespread military deployments to states plagued by cartel violence, a policy that continued under his successor, Enrique Peña Nieto (2012–2018). Over time, however, these deployments led to significant human rights violations, including arbitrary detentions, torture, and extrajudicial killings (Brewer 2009; Human Rights Watch 2024).

The increase in human rights abuses has been exacerbated by the government's strategy of decapitating drug cartels, which destabilized existing structures and led to inter-cartel wars (Osorio 2015; Durían-Martínez 2017). Reports showed a sharp rise in complaints of serious abuses committed by federal security forces during the so-called war on drugs, including more than 5400 civilians killed by the army(Human Rights Watch 2024).

To predict the location of military misconduct in Mexico, I use a unique municipality-level panel dataset with annual measures of human rights complaints filed against the armed forces and federal police from 2000 to 2016. This dataset includes detailed information on 12,437 allegations of serious human rights abuses, such as torture, extrajudicial killings, and illegal detentions, drawn from records of the National Human Rights Commission (CNDH). The data were collected by Flores-Macías and Zarkin (2024) after winning a series of appeals to access government information through Mexico's National Institute for Transparency, Access to Information, and Personal Data Protection (INAI). The CNDH data is particularly valuable as it provides a proxy measure for detailed subnational human rights abuses by federal security forces, given the high levels of public trust and recognition of the CNDH as the principal body for reporting such abuses in Mexico (Valencia 2006).

In addition to data on human rights complaints, I use municipality-level annual panel data collected by several institutional sources[4] and organized into predictive modeling categories. First, socioeconomic and demographic factors include annual variables such as population size and demographics related to age and gender, alongside health, education, income, and social indices. The second category captures the political and institutional context with variables measured annually, such as political alignments between local, state, and national governments. The third category comprises military and security factors, represented by annual measures such as the homicide rate, which serves as a proxy for violence levels, and years with military presence. Finally, geographic and environmental conditions include variables such as the municipality's rural index, soil index, altitude, and water availability that might affect policing strategies. To account for broader temporal and spatial variations, I also include a year variable and state dummies.

Methodology: machine-learning models and super learner ensemble

Machine learning techniques have become a prominent analytical tool in policing and violence studies over the past 35 years (Mastrobuoni 2020). Traditionally, research in this area has focused on evaluating the performance of a single algorithm, typically assessing metrics such as accuracy, sensitivity, and specificity (see, for example, Cubitt and Birch 2021; Cubitt et al. 2022). Departing from this narrower approach, I leverage an ensemble modeling framework that combines the predictive strengths of multiple algorithms to optimize overall performance (Van der Laan et al. 2007; Mayer 2023). This approach builds on successful applications in conflict studies (Bazzi et al. 2022) and investigations into other forms of misconduct (Gallego et al. 2022).

To predict police and military misconduct, I train a diverse set of machine learning models, including Lasso, Random Forests, Extreme Gradient Boosting, and Neural Networks. Each of these models brings distinct advantages and limitations, making ensemble methods particularly valuable for balancing their predictive strengths (Foster et al. 2016). Rather than relying on theoretical assumptions, I allow the data to guide the selection of the best-performing model based on out-of-sample performance.

To ensure robust results, I divide the dataset into training and testing sets. For the training set, I employ a tenfold cross-validation procedure to fine-tune the models and identify optimal parameter combinations. This iterative process partitions the training data into equal-sized subsamples, using each in turn for validation. In the following subsection, I detail the five machine learning methods applied in this study, drawing on foundational works such as Marsland (2011) and Foster et al. (2016), while highlighting foundational and substantive research that has employed these techniques.

Prediction methods. The Lasso regression—similar to a logistic regression model—adds a penalization term based on the sum of the absolute values of the coefficients and a penalization term based on the sum of the square of the parameters. By incorporating these penalization terms, the model parameters are driven toward zero, resulting in a more streamlined and efficient model compared to logistic regression. The tuning parameters in the cross-validation are the weight of the penalization terms in the objective function and the relative weight of the absolute sum of coefficients as the penalization term (Tibshirani 1996). In this way, this algorithm is the simplest of the five I test, and the result is a simple model less prone to overfitting.

Random Forests is a widely used algorithm for predicting misconducts (Cubitt and Birch 2021), known for its superior classification accuracy. The algorithm operates by constructing an ensemble of decision trees, each built from random subsets of both the training data and predictor variables. These trees act as a series of conditional splits, dividing the dataset into distinct groups, or leaves, based on specific variable thresholds. Within each leaf, the predicted outcome corresponds to the most frequently observed result among the training data assigned to that leaf. The final output is derived from a collective consensus, achieved either by averaging the predictions for regression tasks or taking a majority vote for classification problems (Foster et al. 2016). In my approach, the ensemble comprises 500 trees, with the optimal number of features considered at each split determined through cross-validation. This process not only ensures diversity within the forest but also maximizes predictive performance by leveraging the strengths of each tree in the ensemble.

Extreme Gradient Boosting Machines (XGBoost) are ensembles of weak learners, in this case, decision trees. Unlike Random Forests, which operate with independently fitted trees, XGBoost employs boosting, sequentially applying classification algorithms to a reweighted iteration of the training data (Foster et al. 2016). This approach improves model performance by addressing the weaknesses of preceding trees through gradient-based adjustments to the loss function. Each subsequent predictor learns from the errors of its predecessors, refining the model incrementally through a gradient descent procedure to minimize loss (Freund et al. 1999). In my methodology, I maintain a fixed learning rate and a minimum number of observations in terminal nodes to mitigate overfitting. Through the cross-validation procedure, I determine the optimal number of trees and interaction depth, ensuring the model's robustness and predictive accuracy.

Neural networks model the relationship between inputs and outputs in a manner similar to biological brains. These models consist of three fundamental components: an activation function that transforms the weighted sum of inputs (predictors) into an output for each neuron; a network topology, which includes the arrangement of neurons, layers, and their interconnections; and a training algorithm that adjusts the weights of these connections based on the input signals to activate neurons accordingly. This training shapes the model's final predictions. The optimization challenge involves identifying the best weights for the input signals at each node (See Marsland 2011 for more information about neural networks). In my analysis, I maintain a fixed logistic activation function and employ cross-validation to optimize the number of neurons in the hidden layer (size) and the regularization parameter (decay).

Ensembles are aggregates of multiple models that work together to deliver a final prediction. Typically, ensembles outperform their individual model components because they combine different models' strengths. In my analysis, I utilize the Super Learner ensemble method outlined by Mayer (2023). This approach seeks to optimize the blend of individual models by minimizing their cross-validated out-of-bag risk. According to Van der Laan et al., (2007), this type of ensemble model can perform comparably to the best possible weighted combination of its constituent algorithms in the long run.

Findings

In this section, I show the results of the predictive performance of four machine learning models trained to identify instances of police and military misconduct in Colombia and Mexico. The models include a regularized logistic regression model (Lasso), an extreme gradient boosting classifier (XGBoost), a neural network model, and a random forest model (ranger). Each model was assessed using a 10-fold cross-validation procedure, and their performance was measured by the area under the Receiver Operating Characteristic (ROC) curve.

Predictability of extrajudicial killings at the municipality level in Colombia. Figure 1 and Table 1 illustrate the performance of four predictive models, revealing substantial differences in their ability to classify military misconduct in Colombia. Among the models, XGBoost emerges as the strongest performer. It achieves a mean ROC score of 0.88 and a remarkable sensitivity of 0.99, making it highly effective at identifying true misconduct cases. Although its specificity is relatively low at 0.17, XGBoost still performs better than the other models in this regard. This balance between minimizing false negatives and managing false positives makes it the most reliable option for this task.

The Random Forest and Lasso models also exhibit strong sensitivity, successfully detecting most cases of misconduct.

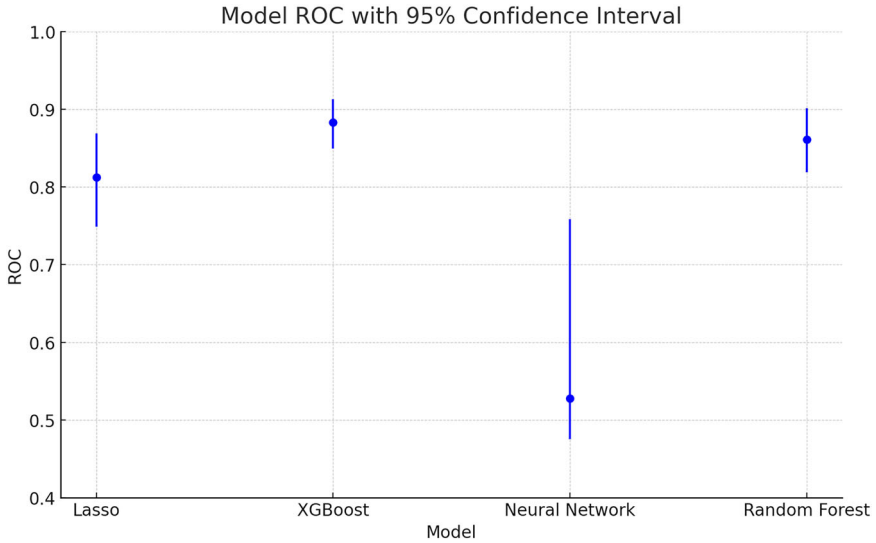

Fig. 1 ROC for different models. Note: The ROC values for each model are displayed with 95% confidence intervals, generated using 10-fold cross-validation.

Table 1 Model performance summary.

Model	ROC	Sensitivity	Specificity
Ranger	0.86	0.99	0.13
Lasso	0.81	0.99	0.01
XGBoost	0.88	0.99	0.17
Neural network	0.50	1.00	0.00
SuperLearner	0.88	0.99	0.18

Note: Model performance is summarized in terms of ROC (Receiver Operating Characteristic curve), sensitivity (true positive rate), and specificity (true negative rate). Values were computed using 10-fold cross-validation on the training data.

However, their specificity falls even lower, further increasing the rate of false positives. Although these models effectively capture true misconduct cases, their inability to reduce misclassification of non-misconduct instances limits their practical applicability. The neural network model performs poorly in this context. Its mean ROC score of 0.5 suggests that it does no better than random guessing. Additionally, it fails to achieve any specificity, consistently misclassifying all cases as positive. This fundamental flaw undermines its ability to distinguish between misconduct and non-misconduct instances, making it unsuitable for predictive modeling in Colombia. Due to its lack of robustness, I excluded the neural network from further analysis and redirected focus toward improving the models with stronger performance.

To enhance predictive accuracy and balance the strengths of individual models, I implemented a meta-ensemble approach. The results of the meta-ensemble show a significant improvement in the model's overall classification ability (see Table 1). The ensemble (Super-Learner) achieved an ROC of 0.8, which is in line with the top-performing individual models. The ensemble's sensitivity remained high at 0.99, meaning that it was able to detect almost all instances of extrajudicial killings. However, the specificity of the ensemble, at 0.18, reflects continued challenges in correctly identifying true negatives, though this still represents an improvement over some individual models like Lasso and Neural Networks.

The coefficient estimates from the ensemble provide insight into the contribution of each base learner to the final ensemble. Notably, the coefficients for both XGBoost and Random Forest are highly significant, suggesting that these models had the strongest influence in the final ensemble predictions. In contrast, the contribution of the neural network model was not statistically significant ($p = 0.625$), indicating that it played a lesser role in the ensemble; hence, I decided to exclude it from the final analysis presented later in this section. The Lasso model also contributed significantly to the ensemble, further confirming its utility in combination with the more complex tree-based methods.

The performance of the models is further illustrated in Fig. 2 which displays the ROC curves for the individual models—Random Forest, Lasso, and XGBoost—as well as the Super-Learner. The ROC curves demonstrate the relationship between sensitivity and specificity for each model across the test data. The diagonal dashed line represents the line of no-discrimination, where the model would perform no better than random guessing. The figure shows that, the ensemble and XGBoost models consistently outperform Lasso and Ranger across a range of specificity values. The ensemble model consistently outperforms the individual models across a range of specificity and sensitivity values, reflecting the ensemble's ability to aggregate the strengths of its individual models.

How to improve specificity?. While the initial models demonstrated strong performance in terms of ROC and sensitivity, their specificity was weaker. Improving specificity is critical for military misconduct because of the potential trade-offs between false positives and false negatives. A false positive—wrongly predicting misconduct where none occurs—may lead to strained relations between oversight bodies and armed forces, and a loss of legitimacy among actors unjustly accused (Brooks and Greenberg 2021). In contrast, a false negative—failing to predict an actual case of misconduct—could have severe human rights implications, undermining public trust and the credibility of state institutions (Curtice 2021). Therefore, from a theoretical perspective, increasing specificity aligns with principles of resource efficiency and institutional legitimacy. While high sensitivity remains a priority to avoid overlooking extrajudicial killings, balancing specificity is essential to minimize the consequences of false positives.

To improve the balance between sensitivity and specificity, I adjusted the classification thresholds for each model. By leveraging ROC curve analysis and employing methods such as Youden's index and the closest-to-top-left approach (following Robin 2011), I identified new thresholds that enhanced the specificity of the models without significantly compromising their

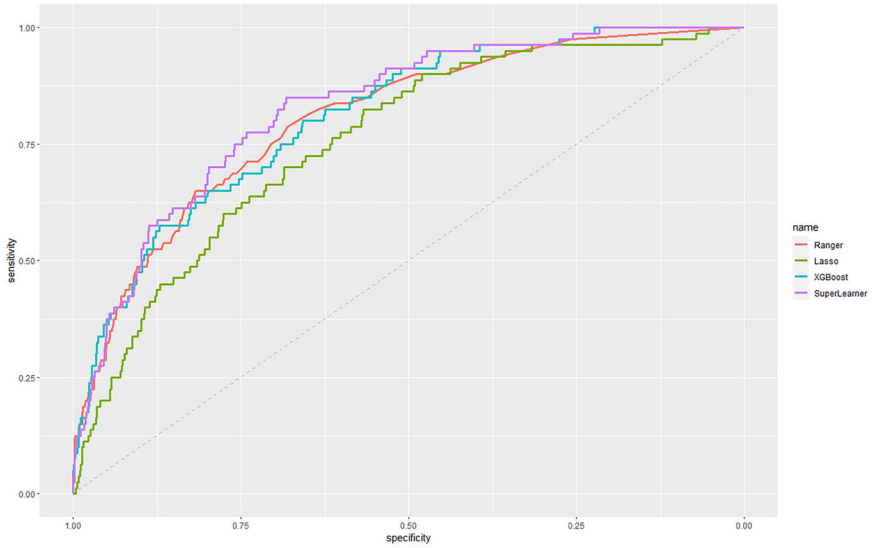

Fig. 2 ROC curve for the municipality-level prediction in Colombia. Note: ROC curve for the municipality-level prediction in Colombia. Each line represents the ROC curves for each of the estimated models.

Table 2 Threshold optimization results.

Model	Optimal threshold	Sensitivity	Specificity
Ranger	0.02	0.68	0.78
Lasso	0.04	0.68	0.70
XGBoost	0.01	0.65	0.80
SuperLearner	0.97	0.85	0.68

Note: These performances were calculated following Robin (2011)'s approach to models' specification. See the Supplementary Appendix for more metrics and details.

Table 3 Model performance summary for Mexico.

Model	ROC	Sensitivity	Specificity
Ranger	0.9214	0.9924	0.2985
Lasso	0.9094	0.9834	0.3664
XGBoost	0.9233	0.9817	0.4502
Neural network	0.8762	0.9932	0.0835
SuperLearner	0.9245	0.9822	0.4530

Note: Model performance is summarized in terms of ROC, sensitivity, and specificity, computed using 10-fold cross-validation on the training data.

sensitivity. The threshold adjustments led to notable improvements in specificity across the models, see Table 2. In particular, tree-based models like Extreme Gradient Boosting and Random Forest showed marked increases in their ability to correctly classify negative cases. Although these adjustments resulted in slight reductions in sensitivity, the overall model accuracy remained competitive. Similarly, the Lasso model benefited from threshold optimization, achieving a better balance between specificity and sensitivity, enhancing its classification performance. Overall, these adjustments resulted in more robust models that effectively trade off sensitivity for specificity to improve performance without sacrificing too much predictive power.

Predictability of human rights violations complaints in Mexico. Table 3 and Fig. 3 summarize the performance of each model used in the Mexican context. The models generally exhibited strong accuracy, with all around 90 percent, although, similarly to the Colombian case, they all struggled with low specificity to varying degrees (from 0.08 to 0.45). Similar to the previous estimation, the XGBoost model emerged as the best performer with the highest mean ROC value of 0.923. It not only demonstrated strong sensitivity but also managed a more balanced specificity at 0.450. This model proved efficient at minimizing both false positives and false negatives, making it highly suitable for balanced classification tasks.

The Lasso and Ranger models displayed notably good ROC across the 10 validation folds, reflecting their effective discrimination between true positive and false positive instances. However, their specificities remained substantially low at 0.36 and 0.29, respectively, underscoring the models' propensity to generate false positives. In contrast, the neural network model lagged significantly behind the others, with a mean ROC of only 0.876 and a problematic specificity of virtually zero, reflecting severe overfitting to positive cases. This makes it unsuitable for scenarios where a balanced prediction of true positives and true negatives is crucial.

To enhance predictive accuracy and integrate the strengths of these individual models, I adopted a meta-ensemble approach. The ensemble model, referred to as SuperLearner, achieved a superior ROC of 0.924. It maintained high sensitivity at approximately 0.982 while improving specificity to 0.453 compared to some individual models like the Lasso and neural network.

The coefficients from the ensemble model elucidate the contribution of each base model to the final predictions. Notably, the coefficients for XGBoost and Ranger were highly significant, indicating that these models had the strongest influence on the ensemble's predictions. In contrast, the contribution of the neural network was minimal and not statistically significant, leading to its exclusion from the final analysis. This reflects the ensemble's capacity to leverage the strengths of the more effective models.

Figure 4 presents the ROC curves for each of the four models and the ensemble for municipality-level predictions in Mexico. Better than the Colombian case, the curves for all models are far

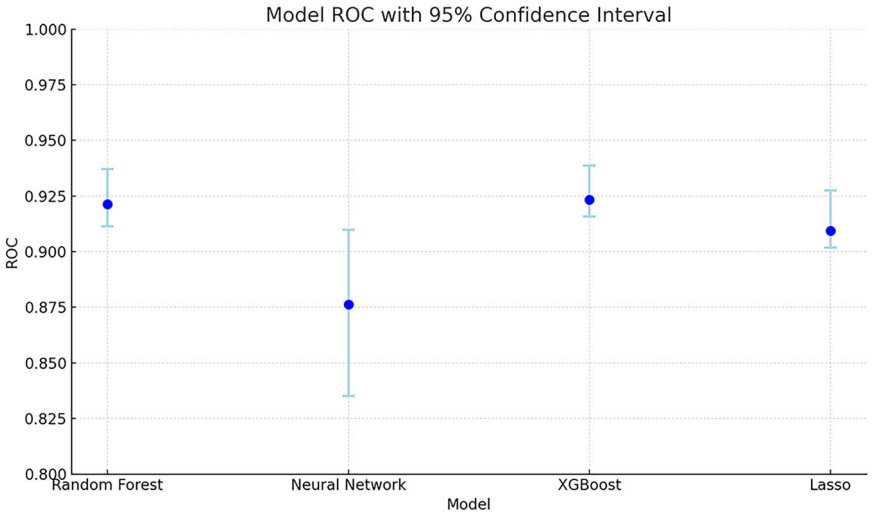

Fig. 3 ROC for different models in Mexico. Note: The ROC values for each model, presented with 95% confidence intervals, were generated using 10-fold cross-validation.

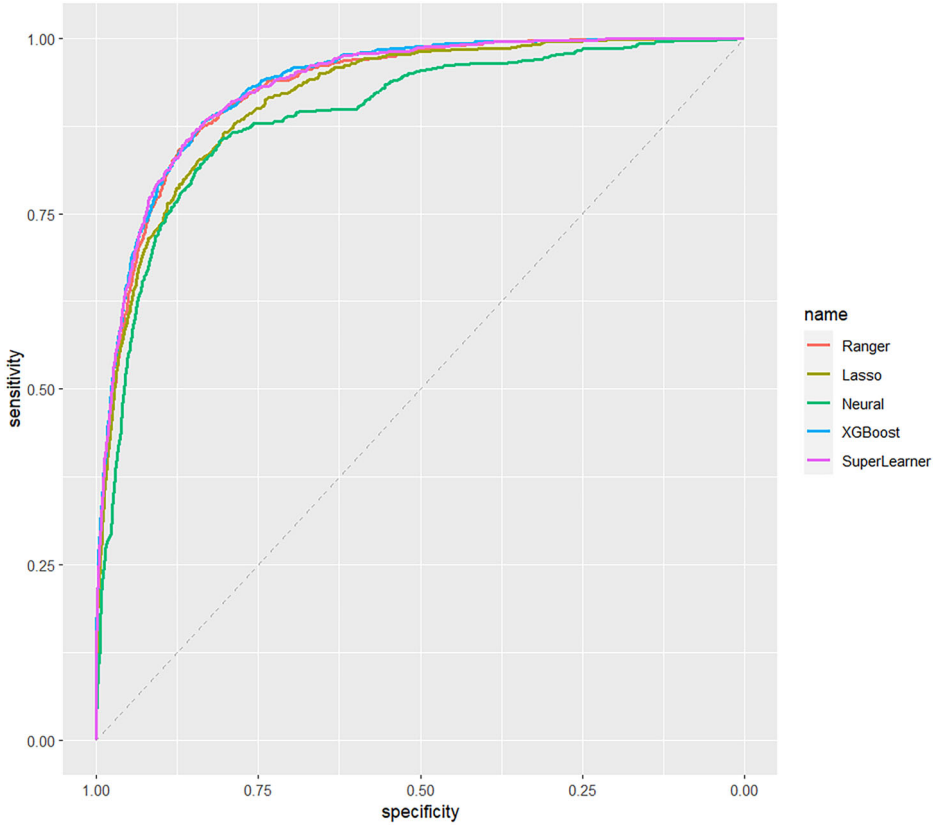

Fig. 4 ROC curves for the municipality-level prediction in Mexico. Note: ROC curve for the municipality-level prediction in Mexico. Each line represents the ROC curves for each of the estimated models.

from the 45° line, indicating that all classifiers perform better than a naive model. Notably, the Random Forest (Ranger) and Extreme Gradient Boosting (XGBoost) demonstrate superior performance, achieving the highest ROC scores. The ROC curve for the ensemble closely follows that of the top-performing models and, under certain levels, outperforms them. The area under the ROC curve (AUC) for the ensemble reaches approximately 0.924, reinforcing its enhanced classification performance compared to the individual models.

How to improve specificity?. Similar to the Colombian case, while the models deployed in the Mexican context initially displayed commendable performance in terms of ROC and sensitivity, their specificity was relatively low, leading to a higher rate of false positives[5] To address this imbalance and enhance the models' specificity without significantly impacting their sensitivity, I adjusted the classification thresholds for each model. This adjustment was guided by ROC curve analysis and employed optimization techniques such as Youden's index and the closest-

to-top-left approach (following Robin 2011), which helped in identifying new thresholds that better distinguished between positive and negative instances.

These threshold adjustments resulted in significant improvements in specificity across all models, see Table 4. All the models experienced an increase to a more balanced level when the new thresholds were applied. These adjustments slightly reduced their sensitivity but substantially decreased the false positive rates, enhancing their overall utility in practical settings. This fine-tuning allowed for better overall classification performance, striking a desirable balance that optimizes the trade-off between sensitivity and specificity, thereby improving the reliability of the models for practical deployment.

Table 4 Threshold optimization results for Mexican models.

Model	Optimal threshold	Sensitivity	Specificity
Ranger	0.1338	87.22%	84.15%
Lasso	0.1108	84.39%	82.43%
Neural Network	0.0489	81.02%	85.34%
XGBoost	0.0946	84.61%	86.92%
SuperLearner	0.9474	87.05%	84.58%

Note: Thresholds were optimized using either "youden" and "closest.topleft" best methods.

Feature-importance analysis

To better understand the contributions of different variables in predicting extrajudicial killings in Colombia and human rights violations in Mexico, I calculated the feature importance for the models that performed best in the analysis (the SuperLeaner and XGBoost models). Although this analysis is predictive and not designed to establish causal relationships, feature importance helps identify which factors most strongly influence the model's outcomes by highlighting the relative impact of each predictor variable.

As shown in Fig. 5, Geographic and Environmental Factors emerge as the most significant predictors of extrajudicial killings in Colombia, suggesting that location-based characteristics such as rurality, size of the municipality, and proximity to key resources or capitals heavily influence the likelihood of state violence. These variables likely serve as proxies for state presence and accessibility (Bazzi et al. 2022), with more isolated or difficult-to-reach areas experiencing higher levels of military misbehavior. Future research might investigate these findings to causally test how logistical and environmental challenges can lead to higher coercive state practices.

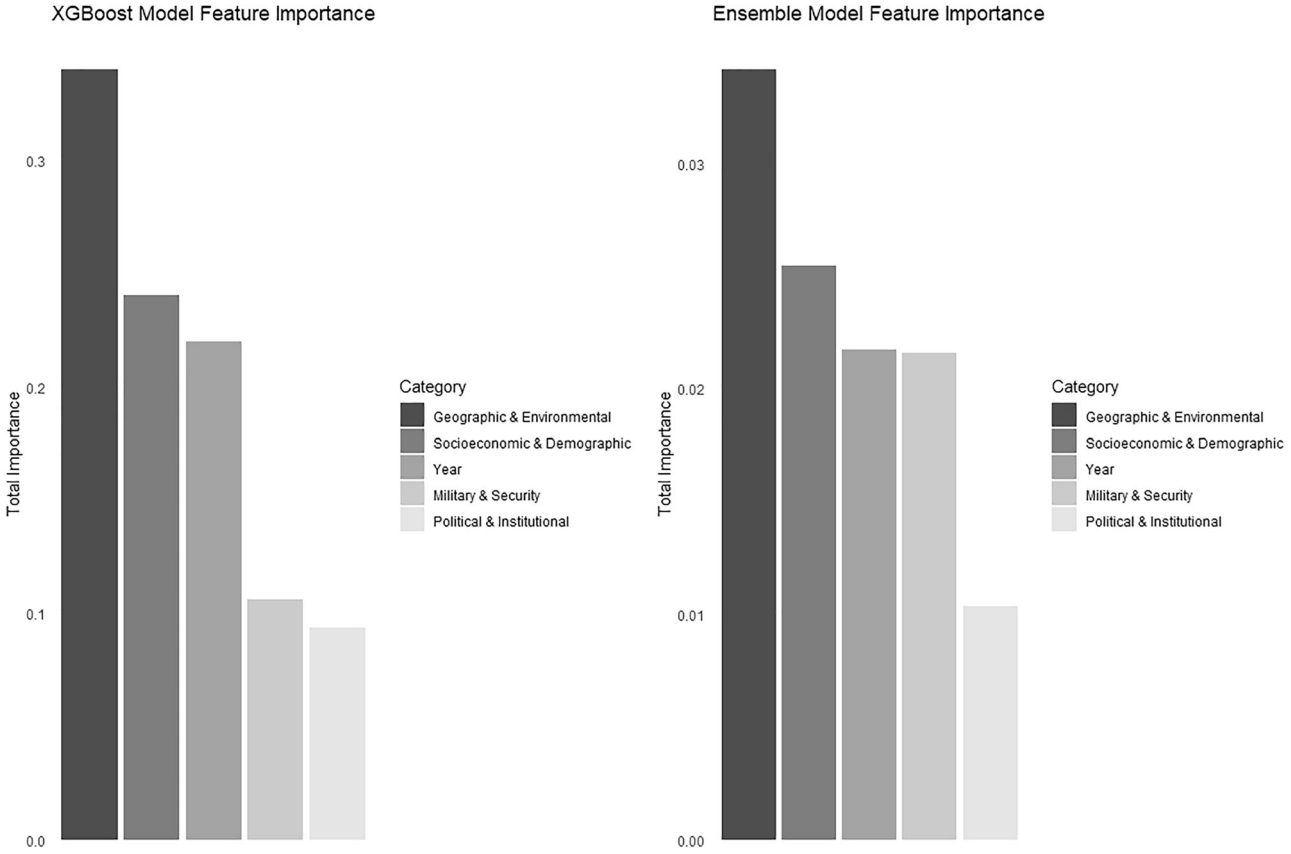

Fig. 5 Feature importance analysis for extrajudicial killings in Colombia using XGBoost and Ensemble models. Note: Categories are grouped as follows. Geographic and Environmental Factors: rurality, altitude, rainfall, water availability, proximity to Bogota, size of the municipality, distance to the principal regional market and the capital city of the department, soil quality, and the region of the municipality (Caribbean, Pacific, Orinoquía, Amazon, or Andean region); Socioeconomic and Demographic Factors: standardized tests (mathematics, language, sciences), population size, Unsatisfied basic needs, tax collection, unemployment, historical and current presence of minorities, and rural index; Year: captures temporal variations; Military and Security Factors: military presence, coca cultivation, conflict intensity, historical conflict, guerrilla, paramilitary, and state attacks, infantry troops, and the rank of the commander (colonel or general); Political and Institutional Factors: judicial capacity, social mobilization, and the presence of churches.

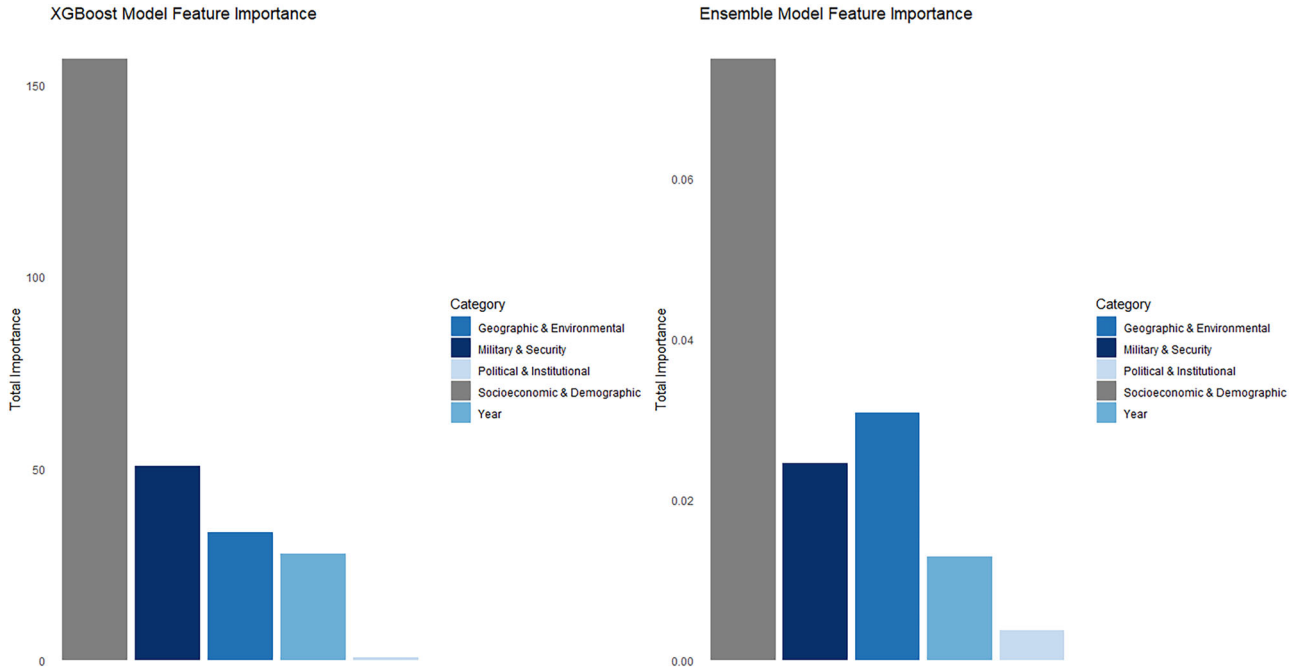

Fig. 6 Feature importance analysis for extrajudicial killings in Mexico using XGBoost and Ensemble models. Note: Categories are grouped as follows. Socioeconomic and Demographic Factors: Log of population size, marginality index, percentage of population aged 15–29, percentage of population aged 15–44, health, education, income and marginality index. Military and Security Factors: Military operations in the area, years of military operations, and homicide rate. Geographic and Environmental Factors: Rural indicator, soil quality, altitude, water availability, and dummy for each State. Political and Institutional Factors: Alignment of local government with presidential party, and governor's alignment with the president. Year: Captures temporal variations and trends over the study period.

Likewise, Socioeconomic and Demographic Factors also weigh heavily in the Colombian models, indicating that regions with lower levels of social development (Unsatisfied Basic Needs), unemployment, and educational disparities are more prone to state violence. These variables reflect broader structural conditions that may exacerbate social tensions and contribute to the potential for human rights violations. Military and Security Factors, such as military presence and coca cultivation—often assumed to be primary drivers of state violence—play a surprisingly smaller role. This suggests that the presence of armed conflict or security forces alone does not fully account for the distribution of extrajudicial killings, pointing instead to a more complex interplay between geography, social vulnerability, and state action. Political and Institutional Factors, like judicial capacity and social mobilization, have the least predictive power, perhaps reflecting the limited direct influence of institutional quality on immediate security outcomes.

Based on Fig. 6, Socioeconomic and Demographic Factors emerge as the predominant category influencing extrajudicial killings in Mexico across both the XGBoost and Ensemble models. This might suggest that the impact that elements such as health, education, and income disparities have on the likelihood of state violence. These factors may act as indicators of broader societal stressors that, when combined with security measures, could escalate into human rights abuses.

In the XGBoost model, Military and Security Factors follow closely, signifying the significant role of military operations and local violence levels. However, in the Ensemble model, Geographic and Environmental Factors take precedence over military aspects. This variance suggests that while the presence and activities of security forces are relevant, the physical and environmental context of different regions also might shape the interactions between the state and its citizens, particularly in terms of accessibility and logistical challenges that may affect the state's ability to govern effectively. Different from the Colombian case, yearly trends appear to play a limited role, indicating that the structural conditions leading to extrajudicial killings remain stable across the time frame analyzed.

Discussion and limitations

Measuring and predicting human rights violations is inherently challenging due to their clandestine nature, which often leaves them unobserved and dependent on victims or witnesses to report them. Therefore, as with most of the predictive models, this research faces limitations, particularly regarding data quality. A key concern is that predictions may reflect where misconduct is detected rather than where it actually occurs, raising important questions about data reliability. For both Colombia and Mexico, I used proxies to measure police and military misconduct based on efforts by researchers to document armed forces violence. In Colombia, Acemoglu et al. (2020) rely on data from a Jesuit NGO that aggregates direct reports from the field, clergy accounts, and detailed media analysis. In Mexico, Flores-Macías and Zarkin (2024) employ official data from the National Human Rights Commission (CNDH), which is widely recognized for its credibility in Mexico. While these proxies are among the best available, they inherently reflect limitations in capturing the full extent of human rights violations from the armed forces.

In addition to data quality limitations, the use of artificial intelligence in law enforcement raises ethical concerns. Although the prediction methods applied in this research represent cutting-edge approaches and are well-established in the literature, specific challenges arise when predicting policing outcomes (Berk 2021). Methods like those employed in this paper (e.g., Random Forest or XGBoost) have been associated with risks of surveillance overreach and discrimination against marginalized groups (see, for example, Karppi (2018), for a discussion about ethical

concerns in predictive policing). This research, in particular, aims to prevent human rights violations by armed forces rather than identifying individuals as culpable before such violations occur. To achieve this goal, the implementation of these methods must be accompanied by robust human oversight and governance mechanisms. These measures are necessary to monitor and address potential biases, ensuring that such tools enhance accountability and do not exacerbate existing systemic issues.

A further consideration is the operational context of the armed forces in Latin America. I have used the term Armed Forces to refer to both the police and the military. While these institutions have distinct legal mandates, structures, and functions in many countries, Latin America presents a unique case where such distinctions have increasingly blurred. Over the past decades, the military has been systematically deployed for internal security and law enforcement duties, a process often described as constabularization or militarization of public security (Flores-Macías and Zarkin 2021). This shift has led to the military assuming roles traditionally reserved for police forces, such as patrolling urban centers, conducting counter-narcotics operations, and engaging in direct law enforcement activities (Pion-Berlin 2017; Blair and Weintraub 2023). While institutional differences remain, particularly in training, doctrine, and oversight mechanisms, the functional reality in both Mexico and Colombia supports the use of armed forces as an umbrella term in this research. In both contexts, the term armed forces is appropriate because, functionally, both military and federal police forces have been involved in security operations that would typically fall under the purview of either institution.

Conclusion

This research shows the potential of machine learning models to predict military and police misconduct, using data from extrajudicial killings in Colombia (2000–2010) and human rights violations in Mexico (2000–2016). By training models such as Lasso, Random Forests, Extreme Gradient Boosting, and Neural Networks on subnational data, I show that misconduct can be reliably anticipated at the municipality level. Ensemble methods, particularly the SuperLearner, further enhanced predictive accuracy, balancing sensitivity and specificity to identify areas and contexts at high risk for abuses. The feature importance analysis revealed that, in Colombia, Geographic and Environmental Factors, and in Mexico the Socioeconomic and Demographic variables, play the most significant role in predicting these abuses.

The findings carry several important policy implications. First, the ability to predict misconduct at granular levels presents an opportunity for governments to implement proactive measures, such as Early Intervention Systems (EIS) that flag at risk officers and municipalities. By incorporating geographic and socioeconomic data, these systems could prevent misconduct by focusing resources on vulnerable areas before violations occur (see, as an example, Carton et al. 2016). Moreover, the ability to detect potential misconduct before it happens aligns with ongoing efforts in Latin American countries and beyond to reform security institutions and reduce human rights violations (González 2020). While this study focuses on Colombia and Mexico, the methods and insights are broadly applicable to other contexts as presented above. Predictive models that leverage subnational data can be adapted to different institutional and cultural settings to identify patterns of misconduct and inform targeted interventions. By doing so, governments worldwide could mitigate both the human and institutional costs of abuses, fostering greater public trust and accountability in security institutions.

Building on these implications, we might consider how Armed Forces misconduct differs between the strategic and operational-tactical levels and how prevention efforts must account for these distinctions. At the strategic level, predictive tools can inform high-level decisions, such as resource allocation, deployment strategies, and institutional reforms, ensuring that policies prioritize accountability and ethical standards (Khalifa 2021). By integrating data-driven insights into strategic planning, governments can address systemic conditions that enable misconduct, such as flawed incentives or insufficient oversight mechanisms. At the operational-tactical level, these tools can guide interventions in high-risk areas, such as targeted training programs or enhanced field supervision, to mitigate situational pressures that lead to misconduct. Aligning these efforts across both levels not only enhances their effectiveness but also ensures that tactical decisions reinforce broader strategic objectives, creating a cohesive framework to reduce human rights violations and strengthen public trust in security institutions.

Data availability

Users can access the data and review the R script for analysis in the following replication package available on Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GBZJ9O.

Received: 20 November 2024; Accepted: 28 April 2025;
Published online: 05 June 2025

Notes

1 I acknowledge the significant distinctions between police forces and the military, including their legal mandates, structures, and functions. However, for the sake of brevity, I use the term "armed forces" to refer to both institutions, police and military, and may use them interchangeably throughout this paper.
2 Predictive models like this one are increasingly recognized as valuable tools in criminal justice and policing, where they are used to anticipate crimes or identify high-risk areas. For example, Berk (2017) has highlighted the potential of machine learning for risk assessment in criminal justice, such as predicting future offending or identifying individuals at risk of reoffending. The strength of these models lies in their ability to inform proactive measures and resource allocation, particularly in contexts where causal inference methods may not be feasible or appropriate.
3 I also used variables published by Acemoglu et al. (2020), Ahmed et al. (2021), and Gelvez and Johnson (2023). See the Supplementary Appendix for data description.
4 This dataset includes information from the National Institute of Statistics and Geography (INEGI), the United Nations, and data collected previously by other researchers such as Ley (2018) and Angulo (2023). See the Supplementary Appendix for more information.
5 In the Colombian case above, I discussed the theoretical trade-offs between false positives and false negatives, emphasizing the implications of these outcomes for institutional legitimacy and trust between oversight bodies and the armed forces.

References

Acemoglu D, Fergusson L, Robinson JA, Romero D, Vargas J (2020) The perils of high-powered incentives: Evidence from Colombia's false positives. Am Econ J Econ Policy 12(3):1–43
Ahmed AT, Johnson M, Vásquez-Cortès M et al. (2021) Slavery, elections and political affiliations in Colombia. J Historical Political Econ 1(3):283–318
Albarracín J, Milanese JP, Valencia IH, Wolff J (2023) Local competitive authoritarianism and post-conflict violence. An analysis of the assassination of social leaders in Colombia. Int Interact 49(2):237–267
Alston P (2010) Report of the Special Rapporteur on Extrajudicial, Summary Or Arbitrary Executions, Philip Alston: Addendum: Mission to the Democratic Republic of the Congo. UN
Angulo JC (2023) Green gold: Avocado production and conflict in Mexico. Working paper, Agricultural and Applied Economics Association (AAEA), Washington D.C, USA. https://ageconsearch.umn.edu/record/335896
Apergis N, Cooray A (2020) How do human rights violations affect poverty and income distribution? Int Econ 161:56–65

Aranguren Romero JP, Cardona Santofimio JN, Agudelo Hernández JÁ (2021) Inhabiting mourning: Spectral figures in cases of extrajudicial executions (false positives) in Colombia. Bull Lat Am Res 40(1):6–20
Bazzi S, Blair RA, Blattman C, Dube O, Gudgeon M, Peck R (2022) The promise and pitfalls of conflict prediction: evidence from Colombia and Indonesia. Rev Econ Stat 104(4):764–779
Bedi M (2015) Unraveling unlawful command influence. Wash UL Rev 93:1401
Berk R (2017) An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. J Exp Criminol 13:193–216
Berk RA (2021) Artificial intelligence, predictive policing, and risk assessment for law enforcement. Annu Rev Criminol 4(1):209–237
Blair R, Karim S, Gilligan MJ, Beardsley KC (2022) Policing ethnicity: Lab-in-the-field evidence on discrimination, cooperation, and ethnic balancing in the Liberian National Police. Q J Political Sci 17:141–181
Blair RA, Weintraub M (2023) Little evidence that military policing reduces crime or improves human security. Nat Hum Behav 7(6):861–873
Brewer SE (2009) Structural human rights violations: The true face of Mexico's war on crime. Hum Rights Brief 16(2):2
Brooks SK, Greenberg N (2021) Psychological impact of being wrongfully accused of criminal offences: A systematic literature review. Med, Sci Law 61(1):44–54
Caforio G (2014) Psychological problems and stress faced by soldiers who operate in asymmetric warfare environments: Experiences in the field. J Def Resour Manag 5(2):23–42
Campbell F, Valera P (2020) The only thing new is the cameras": A study of us college students' perceptions of police violence on social media. J Black Stud 51(7):654–670
Carton S, Helsby J, Joseph K, Mahmud A, Park Y, Walsh J, Ghani R (2016) Identifying police officers at risk of adverse events. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp. 67–76
Cubitt TI, Birch P (2021) A machine learning analysis of misconduct in the New York police department. Polic Int J 44(5):800–817
Cubitt TI, Gaub JE, Holtfreter K (2022) Gender differences in serious police misconduct: A machine-learning analysis of the New York police department (nypd). J Crim Justice 82:101976
Curtice T (2021) How repression affects public perceptions of police: evidence from a natural experiment in Uganda. J Confl Resolut 65(10):1680–1708
Durían-Martínez A (2017) The politics of drug violence: Criminals, cops and politicians in Colombia and Mexico. Oxford University Press, New York
Elbogen EB, Johnson SC, Wagner HR, Sullivan C, Taft CT, Beckham JC (2014) Violent behaviour and post-traumatic stress disorder in us Iraq and Afghanistan veterans. Br J Psychiatry 204(5):368–375
Evans O, Kelikume I (2019) The impact of poverty, unemployment, inequality, corruption and poor governance on Niger delta militancy, Boko haram terrorism and Fulani herdsmen attacks in Nigeria. Int J Manag Econ Social Sci 8(2):58–80
Flores-Macías G, Zarkin J (2024) The consequences of militarized policing for human rights: evidence from Mexico. Comp Political Stud 57(3):387–418
Flores-Macías GA, Zarkin J (2021) The militarization of law enforcement: Evidence from Latin America. Perspect Politics 19(2):519–538
Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J (2016) Big data and social science: a practical guide to methods and tools. Chapman and Hall/CRC, Boca Raton
Franc R, Pavlović T (2023) Inequality and radicalisation: Systematic review of quantitative studies. Terrorism Political Violence 35(4):785–810
Freund Y, Schapire R, Abe N (1999) A short introduction to boosting. Jpn Soc Artif Intell 14(771-780):1612
Gallego J, Prem M, Vargas JF (2022) Predicting politicians' misconduct: evidence from Colombia. In: Data & Policy, ed. 41. Cambridge University Press
Garcia-Sanchez M, Rodríguez-Raga JC (2019) Personality and an internal enemy: understanding the popularity of Álvaro Uribe, 2002-2010. Revista Latinoamericana de Opinión Pública 8(2):89
Gelvez J, Mantilla C, Nieto M (2022) Rewards, sanctions, and trust in the police: Evidence from Colombia. OSF Working Paper. https://doi.org/10.31219/osf.io/aebxy
Gelvez JD, Johnson M (2023) "Los nadies y las nadies": The effect of peacebuilding on political behavior in Colombia. Latin American Politics and Society, Cambridge, United Kingdom, pp. 24–51. https://doi.org/10.1017/lap.2023.34
Gingerich DW, Oliveros V (2018) Police violence and the underreporting of crime. Econ Politics 30(1):78–105
González YM (2020) Authoritarian police in democracy: Contested security in Latin America. Cambridge University Press, Cambridge, United Kingdom
Gordon E (2017) Crimes of the powerful in conflict-affected environments: False positives, transitional justice and the prospects for peace in Colombia. State Crime J 6:132
Greitens SC (2016) Introduction. in dictators and their secret police: Coercive institutions and state violence. Cambridge University Press, Cambridge, pp. 1–20
Hu S, Conrad CR (2020) Monitoring via the courts: Judicial oversight and police violence in India. Int Stud Q 64(3):699–709
Human Rights Watch (2024) World report 2024: Mexico. Accessed: September 24, 2024. https://www.hrw.org/world-report/2024/country-chapters/mexico
Jenasamanta A, Mohapatra S (2022) An automated system for the assessment and grading of adolescent delinquency using a machine learning-based soft voting framework. Humanit Soc Sci Commun 9(1):1–11
Kalyvas SN (2006) The Logic of Violence in Civil War. Cambridge University Press, Cambridge, UK
Kane RJ (2002) The social ecology of police misconduct. Criminology 40(4):867–896
Kappeler VE, Sluder RD, Alpert GP (1998) Forces of Deviance: Understanding the Dark Side of Policing. Waveland Press, Long Grove, Illinois. Accessed: 2025-01-08
Karppi T (2018) "The computer said so": On the ethics, effectiveness, and cultural techniques of predictive policing. Soc Media Soc 4(2):2056305118768296
Khalifa AS (2021) Strategy and what it means to be strategic: redefining strategic, operational, and tactical decisions. J Strategy Manag 14(4):381–396
Knoester M (1998) War in Colombia. Soc Justice 25(72):85–109
Kutnjak Ivković S (2005) Police (mis) behavior: a cross-cultural study of corruption seriousness Polic Int J Police Strateg Manag 28(3):546–566
Lawrence AK (2017) Repression and activism among the Arab spring's first movers: Evidence from Morocco's February 20th movement. Br J Political Sci 47(3):699–718
Lee H, Lim H, Moore DD, Kim J (2013) How police organizational structure correlates with frontline officers' attitudes toward corruption: A multilevel model. Police Pract Res 14(5):386–401
Ley S (2018) To vote or not to vote: how criminal violence shapes electoral participation. J Confl Resolut 62(9):1963–1990
Low HQ, Keikhosrokiani P, Pourya Asl M (2024) Decoding violence against women: analysing harassment in middle eastern literature with machine learning and sentiment anal- ysis. Humanit Soc Sci Commun 11(1):1–18
Marsland S (2011) Learning with Trees and Decision by Committee: Ensemble Learning. Chapman and Hall/CRC, New York, USA
Mastrobuoni G (2020) Crime is terribly revealing: Information technology and police productivity. Rev Economic Stud 87(6):2727–2753
Masullo J, Morisi D (2023) The human costs of the war on drugs: Attitudes towards militarization of security in Mexico. Comp Political Stud 57(6):875–901
Mayer Z (2023) A brief introduction to caretensemble. https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html. Accessed: May-5-2025
Nägel C, Nivette A(2023) Protest policing and public perceptions of police. Evidence from a natural experiment in Germany Polic Soc 33(1):64–80
Ortiz-Ayala A (2021) They see us like the enemy: soldiers' narratives of forced eradication of illegal crops in Colombia. Confl, Security Dev 21(5):593–614
Osorio J (2015) The contagion of drug violence: spatiotemporal dynamics of the Mexican war on drugs. J Confl Resolut 59(8):1403–1432
Pion-Berlin D (2017) A tale of two missions: Mexican military police patrols versus high-value targeted operations. Armed Forces Soc 43(1):53–71
Pridemore WA (2011) Poverty matters: A reassessment of the inequality–homicide relation- ship in cross-national studies. Br J Criminol 51(5):739–772
Robin XT (2011) proc: An open-source package for r and s+ to analyze and compare roc curves. BMC Bioinforma 7:77
Rodríguez Gómez JC (2020) Antecedentes históricos sobre los "falsos positivos" en Colombia
Rowe P (2008) Military misconduct during international armed operations: 'bad apples' or systemic failure? J Confl Security Law 13(2):165–189
Rozema K, Schanzenbach M (2023) Does discipline decrease police misconduct? evidence from Chicago civilian allegations. Am Econ J Appl Econ 15(3):80–116
Salazar-Tobar F, Rengifo AF (2023) Trust in the police in Latin America: A multilevel analysis of institutional and experiential models. Policing J Policy Pract 17:paac113
Stone CE, Ward HH (2000) Democratic policing: A framework for action. Polic Soc Int J 10(1):11–45
Stroud A (2020) Guns don't kill people… : good guys and the legitimization of gun violence. Humanit Soc Sci Commun 7(1):1–7
Sung H-E, Capellan J, Barthuly B (2022) Trust in the police and the militarisation of law enforcement in Latin America. Int J Res Policy 32:311–340
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58(1):267–288
Trust Commission (2024) Los falsos positivos: 6,402 civiles asesinados en estado de indefensi´on
Valencia AG (2006) La justicia social como fin primordial de los derechos humanos. Univ J Aut´onoma de Tabasco

Van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Stat Appl Genet Mol Biol 6(1):1–21
Visconti G (2019) Policy preferences after crime victimization: Panel and survey evidence from Latin America. Br J Political Sci 49(3):1–15
Woolfolk RL, Hannah ST, Wasserman R, Doris JM (2021) Attributions of responsibility for military misconduct: constraint, identification, and severity. Mil Psychol 33(1):1–14

Acknowledgements

I am grateful to Isabella Alcañiz, Matilde Angarita, Juan Carlos Angulo, Brian Kim, the Political Methodology field at the Government and Politics Department of the University of Maryland-College Park, and the attendees of the Future of Science Initiative at the University of Mannheim, Germany, for their insightful comments and support during this research.

Author contributions

Juan David Gelvez wrote the manuscript, performed data analysis and prepared all the figures.

Competing interests

The author declares no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by the author.

Informed consent

This article does not contain any studies with human participants performed by the author.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04967-w.

Correspondence and requests for materials should be addressed to Juan David Gelvez.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025