Journal of Criminal Justice 105 (2026) 102697

Contents lists available at ScienceDirect

Journal of Criminal Justice

journal homepage: www.elsevier.com/locate/jcrimjus

Machine learning for violence prediction: A systematic review and critical appraisal

Stefaniya Kozhevnikova [a], Denis Yukhnenko [a], Giulio Scola [a], Seena Fazel [a,b,*]

[a] Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[b] Oxford Health NHS Foundation Trust, Oxford, United Kingdom

A B S T R A C T

Purpose: To conduct a systematic review of machine learning models for predicting violent behaviour and appraising the validity, usefulness and performance of these models.
Methods: We systematically searched nine bibliographic databases and Google Scholar up to September 2025 for development and/or validation studies on machine learning methods for predicting violent behaviour. We extracted discrimination and calibration performance statistics and evaluated study quality by examining risk of bias and clinical utility.
Results: We identified 38 studies reporting the development and validation of 40 machine learning models. Reporting of performance was mostly limited to the Area Under the Curve (AUC) statistic ($n = 29$, 72%) and around a fifth studies reported calibration performance ($n = 8$, 21%). The range of AUC in the included models was 0.50–0.98. There was a lack of external validation (3 studies, 8%). There was high risk of bias in 31 (82%) investigations, mainly in the analysis domain. There was risk of overfitting due to small samples, lack of transparent reporting, and low generalisability of the models.
Conclusion: Current machine learning models for violence prediction have poor clinical utility. Future work should consider utilising machine learning methods for highly complex data and dynamic predictions for higher precision. Developing more trustworthy models with explainable algorithms and causal predictions should be prioritised.

1. Introduction

Preventing criminal behaviour, and particularly serious crimes involving interpersonal violence, is a core aim of all criminal justice systems. In many low- and lower-middle-income countries, interpersonal violence is one of the leading causes of death or serious injury (United Nations Office on Drugs and Crime, 2023a, 2023b). Even in higher-income countries, such as the USA, domestic violence is a leading cause of maternal mortality (Lawn & Koenen, 2022). Criminal justice is also concerned with preventing reoffending. A meta-analysis of criminal recidivism rates worldwide showed that violent crimes have higher reoffending rates than sexual and traffic-related crimes (Yukhnenko et al., 2023). Further, despite violence being an uncommon outcome in people with mental disorders (Whiting et al., 2021), violence prevention is important to reduce morbidity and costs and enhance recovery, and relevant for other services. These overlapping problems – in criminal justice and its intersection with health and other public services – mean that there is a need for models and tools that can assess and stratify violence risk. Such approaches can inform decision-making about violence prevention measures.

There are many violence prediction tools, also known as risk assessment instruments (Fazel et al., 2022). In recent years, their development has shifted from standard regression-based modelling to more advanced statistical methods, including machine learning. Machine learning (ML) is an umbrella term describing several computational methods that perform a task based on an existing (training) dataset (Alpaydin, 2020). Applying ML methods leads to the development of an algorithm that learns from the patterns and rules of the data, and applies these rules to new input data, without directly programming the task-solving process (Mitchell, 1997). Many different ML methods exist, and new approaches are increasingly applied.

One way to classify ML methods is to divide them into “interpretable” methods and “black box” methods. Interpretable ML methods are algorithms that have established direct connections between predictors and outcomes, such as Naïve Bayesian (John & Langley, 1995) or decision trees (Rokach & Maimon, 2013). Black box methods, on the other hand, are usually more complex algorithms, such as neural networks (Goodfellow et al., 2016) or random forests (Ho, 1998). Therefore, while interpretable methods are more easily understood, black box methods solve tasks in ways that are not intuitive, and reaching an understanding

* Corresponding author at: Department of Psychiatry, University of Oxford, Oxford, United Kingdom.
E-mail address: seena.fazel@psych.ox.ac.uk (S. Fazel).

https://doi.org/10.1016/j.jcrimjus.2026.102697
Received 13 March 2026; Received in revised form 2 July 2026; Accepted 2 July 2026
Available online 11 July 2026

requires additional training in a specific method or in advanced statistics. Black box methods often achieve higher performance than traditional methods (Fajar et al., 2024), including in healthcare (Chen & Asch, 2017); however, a key limitation is that not being able to understand and assess the decision-making process of the tool can create a lack of trust among those using the methods. Thus, the utility of black box models in criminology and medicine remains questionable (Petch et al., 2022). Some experts emphasise the importance of understanding the causes of predictions made by models to save time and resources. A review of clinical prediction models in mental health comes to a similar conclusion: despite rapid progression of the field, clinical utility of existing models remains highly limited (Meehan et al., 2022). Despite these limitations of black-box ML methods, recent methodological advancements and the growing complexity of criminal justice and forensic mental health datasets support their application, particularly for detecting complex, nonlinear patterns that traditional statistical approaches may not capture. ML models have also become increasingly prominent in the research literature making a systematic evaluation of their performance timely. As tools based on statistical approaches such as regression have been comprehensively reviewed elsewhere (Fazel et al., 2022), this review focuses on ML models to address a gap in the existing literature. Recent reviews have either focused on discrimination performance (Parmigiani et al., 2022), or on specific populations and tool evaluations (Ogonah et al., 2023; Ramesh et al., 2018), but a broader review of ML in criminal justice is lacking.

To assess the utility of prognostic (or prediction) models, evaluating them separately from their development process is not sufficient. Current markers of high quality include methods that mitigate against under- or overfitting. Underfitted models do not produce sufficiently accurate predictions. Overfitting, on the other hand, involves model development that uses large amounts of data from the training data, with insufficient outcome events compared to candidate predictors, with consequences that catch noise and idiosyncrasies, instead of estimating new outcomes (Goodfellow et al., 2016). An overfitted model thus fails to generalise patterns beyond the training data, with limited practical value. On the other hand, a well-performing model is able to correctly distinguish between positive and negative outcomes (or discrimination performance), and the produced outcomes are as close as possible to observed results (calibration performance) (Huang et al., 2020; Walsh et al., 2017). To ensure that any prediction model can be implemented in different samples and settings and generalised beyond the training sample, model validation is necessary. While internal validation is typically conducted on a subset of the training set sample, external validation uses data collected either in the different location (geographic) or different time period (temporal) (Alpaydin, 2020).

To address the evidence gaps in the field, we aimed to provide a field synthesis of machine learning studies for violence prediction, to assess performance of these models for different outcomes, to critically assess performance and utility of models, and present recommendations for research. As the focus was on black box approaches, rather than interpretable models, we refer to black box models as ML.

2. Methods

2.1. Protocol and registration

A PROSPERO protocol for this review was prepared before carrying out the initial literature search in July 2024 and is available in Supplementary materials (Supplementary A).

2.2. Literature search

We followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Page et al., 2021), and Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA) (Snell et al., 2023). We searched several databases: Ovid (MEDLINE, Embase, APAPsychArticles, AMED, BIOSIS, Global Health, PsycINFO), dissertation databases, and Web of Science Core Collection to identify the papers reporting the development and/or validation of models predicting any type of violent behaviour (i.e., violent behaviour in general, violent crime, and other). In the later stage, we additionally manually searched Google Scholar and grey literature to identify potentially missed studies. No publication date or language restrictions were set. The initial search was conducted in July 2024 and updated in September 2025. The list of keywords used for all databases is available in Supplementary materials (Supplementary B).

2.3. Eligibility criteria

Studies were considered eligible for inclusion if (i) they described the development, internal and/or external validation of a multivariable black box machine learning model for predicting any violent behaviour in any target sample, (ii) the outcome definition of violence was consistent with the World Health Organization definition: “the intentional use of physical force or power, threatened or actual, against oneself, or against a group or community that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation” (World Health Organization, 2014). We considered for inclusion the following black box methods: neural networks, support vector machines, k-nearest neighbours methods, random forest methods, including gradient boosting, super learners (models using a set of several methods), and natural language processing methods (NLP).

We excluded papers that did not account for violence as a separate outcome, i.e., papers that focus on any form of recidivism. We also excluded studies that predict victimisation of a person, rather than perpetration. In addition, we excluded reviews and theoretical papers. We did not consider those that focused solely on biological variables, such as neuroimaging or genetics, which were outside the scope of this review – we focused on those with potential clinical implementation.

2.4. Study selection

In accordance with PRISMA guidelines, two reviewers were involved in the study selection process. At first, SK reviewed the titles and abstracts of all identified studies, and GS double-screened and validated a random sample of 20% titles and abstracts. The same procedure was used to assess eligible full texts. Any disagreements on study selection were resolved in consultation with SF and DY.

2.5. Data extraction

Data extraction was carried out by SK following the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist (Moons et al., 2014) using the automated template (Fernandez-Felix et al., 2023). Data extraction was later verified by GS. Information on the following variables were collected: (i) sample demographics; (ii) setting in which data were collected, e.g., community, probation, hospital, or prison; (iii) study design; (iv) performance estimates of the model, including measures of discrimination and calibration.

When articles reported development and testing of multiple models, for example, comparing neural network and random forest, we included the model with the highest discrimination and calibration measures. If models assessed different outcomes (i.e., domestic violence and general violence) or used different follow-up periods, we included all outcomes and follow-ups as separate models.

2.6. Summary measures

To assess the predictive performance, we extracted measures of

discrimination and calibration. The area under the receiver operating characteristic curve (AUC) is currently one of the most widely used measures of discrimination, and we extracted it wherever possible. It reports the probability that the model will assign a higher risk score to a randomly chosen violent person than a randomly chosen non-violent person.

The concordance index (c-index) measures a model's ability to discriminate between individuals by correctly ranking them according to their risk. It represents the probability that, for a randomly selected comparable pair of individuals, the individual who experiences the outcome is assigned a higher predicted risk than the individual who does not. For binary outcomes without censoring, the c-index is mathematically equivalent to the area under the receiver operating characteristic curve (AUC) (Hartman et al., 2023).

For AUC and c-index the results can vary between 0.50 and 1, where 1 means perfect discrimination, and 0.50 means that a model discriminates no better than a chance. We did not rely on AUC thresholds because thresholds for prediction model performance should be based on the specific context of the model (Wynants et al., 2019), which cannot be standardised for tools used across different criminal justice and forensic settings. In addition, such thresholds are often chosen post-hoc by the researchers, leading to inflated findings (White et al., 2023). We also extracted other metrics of model performance, such as accuracy or F1 score.

Where possible, we also extracted measures of calibration, such as Brier score or calibration error (CalErr). Since the outcome of interest is binary, we considered root mean square error (RMSE) as calibration measurement, as for binary outcome $RMSE = \sqrt[2]{Brier\ score}$ (Hoessly, 2025).

2.7. Risk of bias and publication bias

The risk of bias within each study was assessed with the Prediction model Risk of Bias Assessment Tool (Moons et al., 2025). The PROBAST was developed specifically for evaluating studies of diagnostic and prognostic prediction models and provides ratings of the risk of bias in four different domains, namely: (i) participants, (ii) predictors, (iii) outcome, and (iv) analysis. PROBAST domains assess how appropriate the procedures carried out during model development and validation were, such as variable selection, internal and external validation and performance measuring. For each domain, there are several guidelines that ensure transparent reporting and minimise the risk of creating under- or overfitting of the model. An important part of the model development that prevents overfitting is using a large enough data set for training, especially ensuring that there is a reasonable ratio of positive outcomes and variables used (EPV). Depending on the model, the EPV ratio should be greater than 10 for regression-based models (Peduzzi et al., 1996) and greater than 20 for models using more complex methods (Austin & Steyerberg, 2017; Moons et al., 2025). Other important procedures to lower the risk of bias include using appropriate handling of missing data (participants should not be excluded due to missing data), avoiding data-driven predictor selection, and using both discrimination and calibration performance measures. Based on the established risk of bias, the overall risk is assessed. The overall risk of bias should be considered high if at least one domain has shown high risk and the model has not undergone external validation, unless the original model was developed with a very large dataset.

2.8. Data synthesis

Due to the high heterogeneity of the reported models, a narrative data synthesis was performed. Narrative data synthesis is the qualitative approach often used for synthesising highly heterogeneous data. It consists of four elements: (i) developing a theory; (ii) a preliminary synthesis; (iii) exploring relationships within and between studies; and (iv) assessing the robustness of the synthesis (Centre for Reviews and Dissemination, 2009). We have used narrative synthesis methods, such as tabulation and data visualisation, to provide the structured description of the included studies.

Among black box methods used, the most common were random forests (12 studies, 30%); 9 studies (22.5%) used gradient boosting methods and 9 neural networks. Three studies (7.5%) reported using super learners – algorithms consisting of several different models. One study used k-nearest neighbours methods, specifically, the unsupervised k-means. Fig. 2 provides a more detailed breakdown of ML methods used.

3. Results

We identified 40 ML models published in 38 studies (Fig. 1). Characteristics of the included studies are presented in Table 1. We divided included studies into two groups: those predicting criminal violence (26 studies) and those focused on (non-criminal) violent behaviour outcomes (12 studies). Within these groups, we created subgroups depending on the outcome. The total number of participants was 2,362,029 people, most were male (87.7%). The list of included studies and results is presented in Table 2.

One study used image recognition of a projective psychological test along with other variables, and 5 (13.2%) investigations used natural language processing to predict outcomes from police reports or medical records. One study (Dobbins et al., 2024) used a model based on the Bidirectional Encoder Representations from Transformers (BERT), an NLP model previously developed by Google (Devlin et al., 2019).

For performance, 29 out of 38 (76%) studies reported the area under the curve (AUC). Nine other studies used accuracy and F1-score as a discrimination metric. Most models reported AUCs in the range of 0.70–0.80 (k = 16, 40%). Three models (7.5%) had AUCs in the range 0.80–0.90, five (12.5%) with AUCs over 0.90, and five models had AUCs below 0.70. Eight studies (21%) reported calibration performance: Brier score (k = 1), RMSE and calibration error (k = 2), calibration plots (k = 5). The numerical value for calibration performance, such as Brier score, RMSE, and CalErr, represents the error rate in the same unit of measure as an outcome, and indicates higher performance the closer it is to zero. Assuming outcomes were binary, calibration performance was similar across all three studies reporting it. All studies reported internal validation, mostly using k-fold cross-validation, and three studies reported conducting external validation (Fig. 3).

Out of 37 studies that were not doctoral dissertations and had more than one author, six (15.8%) had both a mental health or criminology specialist and a computer scientist as either a first, second or senior author. Nine studies were led by computer scientists, and 22 by either a criminologist or a mental health specialist.

Most models were developed on populations from high- or middle-income countries, with high-income countries accounting for the largest share of participants (Fig. 4).

3.1. Risk of bias

Based on PROBAST, risk of bias was high in 31 papers (81.6%). The highest risk of bias was related to the analysis domain (Fig. 5).

In the domain of participant selection, PROBAST tests whether there has been oversampling of participants with positive outcomes and ensures that high-risk participants are not prevalent in the training dataset. We found eight studies (21%) did not provide any assessment of outcome risks and/or adjustment for included cohorts, while the sampling procedures were not described and may have not been randomised.

Most models were developed using the data from retrospective cohort studies, which created a risk of bias in the domain of outcome definition and assessment. In six studies (15.8%), there was a risk that predictor variables could be part of the outcome definition, and in five

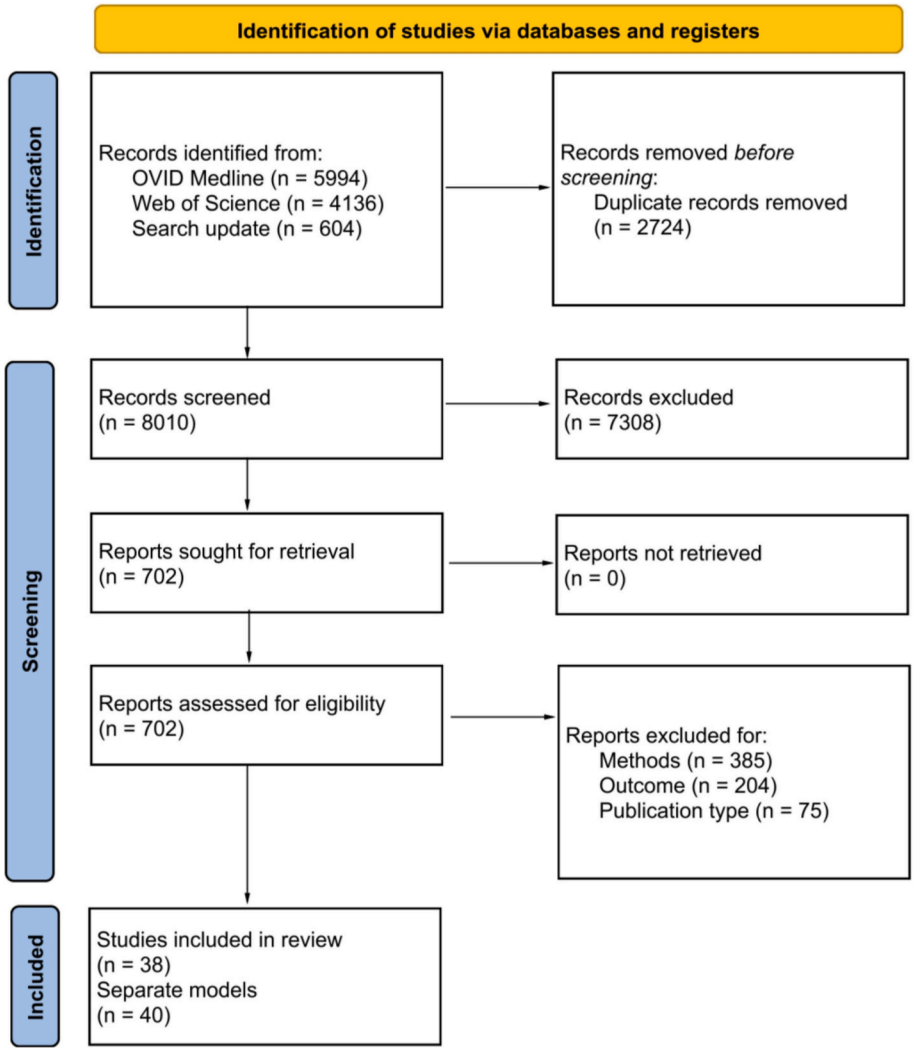

Fig. 1. PRISMA flowchart of included studies. Studies excluded for publication type were conference proceedings, reviews and other reports with no data related to the current review.

Table 1
Features of the included studies divided by the outcomes.

Outcome	Number of studies	Total number of participants with outcome / overall	Mean follow-up time, weeks (years)	Mean participant age, years	Mean male %
		Violent crime			
General criminal violence	6	6424 / 979,428	260.7 (5.0)	34.9	90.5
General violent recidivism	12†	2679 / 583,102	195.6 (3.8)	31.2	88.9
Intimate partner violence	5	398 / 496,248 *	Not reported	Not reported	69.7
Intimate partner violence reoffence	2†	1492 / 71,861	156.43 (2.8)	32.4	100
Extremism	2	1420 / 2328	Not reported	Not reported	Not reported
		Other violent behaviour			
General violent behaviour	4	5251 / 34,460	126.01 (2.4)	32.3	88.5
Inpatient violence	8	2314 / 255,766 *	44.9 (0.9)	38.7	67.4

* – include one or more studies with natural language processing and/or image recognition, where only the overall number of cases is known. † – both outcomes include Zeng et al. (2017) as this study presented models for both outcomes.

studies (13.2%), it is unclear whether outcomes were determined without knowledge of predictor data.

In the analysis domain, events per variable/predictor (EPV) ratio should be over 20, as recommended per PROBAST guidelines. Only six models (15%) had EPV over 20, and another five models (12.5%) had EPV over 10, which would be acceptable for regression-based prediction models. For 9 models, it was not possible to establish EPV, and the other 20 models (50%) had EPV less than 10. Calibration performance was measured for nine models (22.5%) but not reported for two. Moreover, despite a data-driven approach being recommended against for predictive modelling, 8 models (21%) used univariable analysis for the initial predictor selection.

Table 2
All included studies with their results. EPV – events per variable ratio.

Reference	Country	Model	Number of outcomes / number of variables; (EPV)	Discrimination metric	Calibration metric	External validation (performance)
			General criminal violence			
Gordon (1992)	USA	Backpropagation neural network	106 / 17; (6.3)	Recall = 0.76	No	No
Baúto et al. (2024)	Portugal	Multilayer perceptron	59 / 9; (6.6)	Accuracy = 0.83	No	No
Rosellini et al. (2016)	USA	Random forest	5771 / 446; (12.9)	AUC = 0.81	No	Temporal (AUC = 0.77)
Sonnweber et al. (2021)	Switzerland	Stochastic gradient boosting	294 / 10; (29.4)	AUC = 0.76 (95% CI 0.69–0.83)	No	No
Watts et al. (2021)	Canada	XGBoost	863 / 156 (5.5)	Accuracy = 0.57	No	No
Santos et al. (2024)	Brazil	XGBoost	Not reported / 9	AUC = 0.99	No	No
			General violent recidivism			
Grann and Långström (2007)	Sweden	Backpropagation neural network	91 / 10; (9.1)	AUC = 0.72 (95% CI 0.65–0.79)	No	No
Liu et al. (2011)	UK	Multilayer perceptron	343 / 20; (17.1)	AUC = 0.71 (95% CI 0.66–0.75)	No	No
Hamilton et al. (2015)	USA	Neural network	Not reported / 23	AUC = 0.73 (SD 0.03)	RMSE = 0.21 CalErr = 0.05	No
Pelham et al. (2020)	USA	Neural network	Not reported / 43	AUC = 0.78 (95% CI 0.71–0.84)	Brier score = 0.17 (95% CI 0.14–0.19)	No
Neuilly et al. (2011)	USA	Random forest	168 / 60; (2.8)	Accuracy = 0.58 F1 score = 0.62	No	No
Berk and Bleich (2014)	USA	Random forest	1283 / not reported	Accuracy = 0.73	No	No
Salo et al. (2019)	Finland	Random forest	278 / 52; (5.3)	AUC = 0.77 (95% CI 0.70–0.84)	Calibration plot	No
Etzler et al. (2024)	Austria	Random forest	46 / 37; (1.2)	AUC = 0.78 (SD 0.06)	No	No
Zeng et al. (2017)	USA	Stochastic gradient boosting	6455 / 48; (134.5)	AUC = 0.72	Calibration plot	No
Laqueur and Copus (2024)	USA	Super learner	Not reported / 91	AUC = 0.72	Calibration plot	No
Tollenaar and van der Heijden (2013)	Netherlands	Support vector machine	380 / 6; (63.3)	AUC = 0.73	RMSE = 0.40 CalErr = 0.06	No
Wang et al. (2023)	USA	XGBoost (6-months follow-up)	1755 / 41; (42.8)	AUC = 0.85	Calibration plot	Geographic (AUC = 0.69)
		XGBoost (2-year follow-up)	8526 / 41; (207.9)	AUC = 0.78	Calibration plot	Geographic (AUC = 0.68)
			Intimate partner violence			
Karystianis et al. (2021)	Australia	Recurrent neural network; NLP	Not reported	AUC = 0.65	No	No
Hossain et al. (2021)	Bangladesh	Random forest	Not reported / 9	Accuracy = 0.71 F1 score = 0.76	No	No
Garcia-Vergara et al. (2023)	Spain	Random forest	161 / 33; (4.9)	F1 score = 0.87 (SD 0.03)	No	No
Verrey et al. (2023)	UK	Super learner	237 / 13; (18.2)	AUC = 0.71	No	No
Petering et al. (2018)	USA	Support vector machine	Not reported / 26	AUC = 0.69	No	No
			Intimate partner violence reoffence			
Zeng et al. (2017)	USA	Stochastic gradient boosting	1183 / 48; (24.6)	AUC = 0.68	Calibration plot	No
Berk and Sorenson (2020)	USA	Stochastic gradient boosting	144 / not reported	Accuracy = 0.66	No	No
			Extremism			
Ahmed and Okoroafor (2023)	USA	k-Nearest neighbours	180 / 28; (6.4)	Accuracy = 0.91	No	No
Ivaskevics and Haller (2022)	USA	XGBoost	1240 / 79; (15.7)	AUC = 0.87 (SD 0.03)	No	No
			General violent behaviour			
Santamaría-García et al. (2021)	Colombia	Neural network	2117 / 162; (13.1)	AUC = 0.96	Brier score (not reported)	No
Rosellini et al. (2018)	USA	Super learner	829 / 273; (3.0)	AUC = 0.75	No	No
Barzman et al. (2022)	USA	Support vector machine; NLP	54 / 21; (2.6)	AUC = 0.98	No	Geographic (AUC = 0.91)
Kim et al. (2023)	South Korea	Support vector machine; image recognition	134 / not reported	F1-score = 0.98	No	No

(*continued on next page*)

Table 2 (*continued*)

Reference	Country	Model	Number of outcomes / number of variables; (EPV)	Discrimination metric	Calibration metric	External validation (performance)
			Inpatient violence			
Dobbins et al. (2024)	USA	BERT; NLP	246 / not reported	AUC = 0.78	No	No
Menger et al. (2018)	Netherlands	Recurrent neural network; NLP	1248 / not reported	AUC = 0.79 (SD 0.02)	No	No
Bokhari and Hubert (2018)	USA	Random forest	176 / 31; (5.7)	AUC = 0.75	No	No
Wang et al. (2020)	Canada	Random forest	103 / 28 (3.7)	AUC = 0.63 (SD 0.00)	No	No
Cheng et al. (2023)	China	Random forest	611 / 19; (32.2)	AUC = 0.96 (95% CI 0.94–0.97)	No	No
Watts et al. (2023)	Canada	Random forest	26 / 67; (0.4)	AUC = 0.91 (95% CI 0.87–0.95)	No	No
Hu et al. (2023)	Taiwan	Random forest; NLP	406 / not reported	AUC = 0.69	No	No
Le et al. (2018)	Australia	Support vector machine; NLP	Not reported	Accuracy range (0.69–0.77)	RMSE (not reported)	No

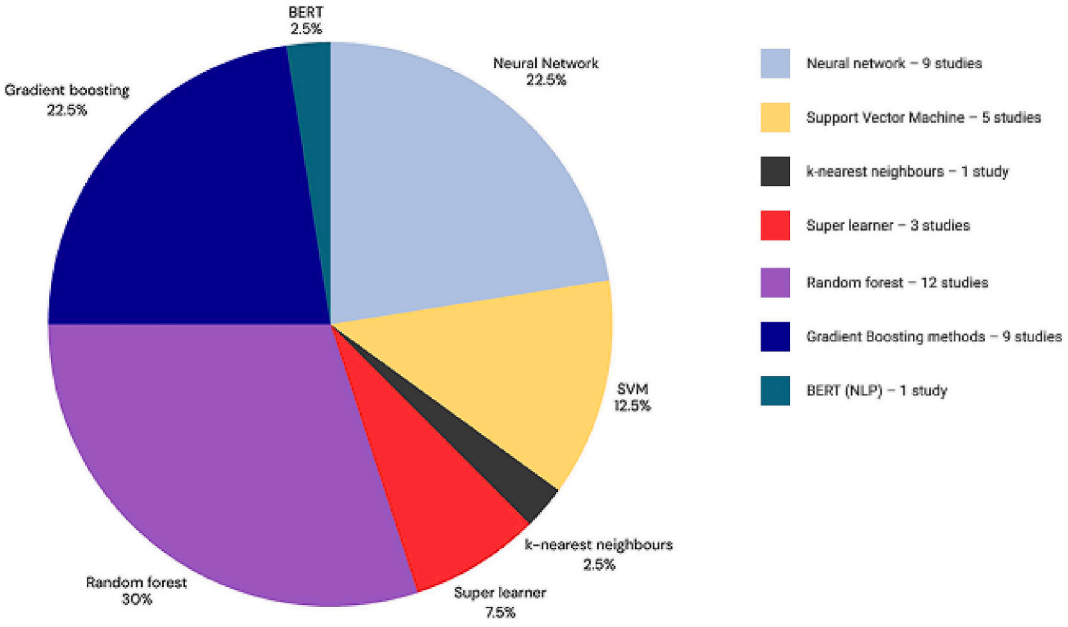

Fig. 2. Distribution of ML methods used in violence prediction studies.

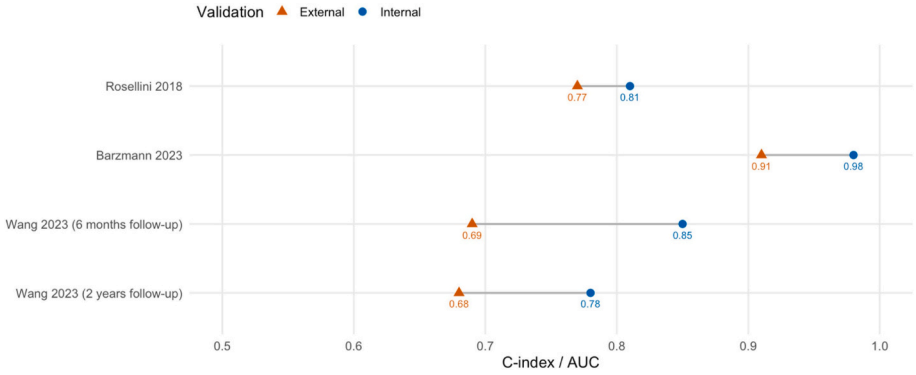

Fig. 3. Discrimination metrics in the studies reporting external validation. Internal AUC reflects model performance on the training dataset, while external AUC reflects performance on a different dataset (e.g., another site or time period). AUC ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination).

Overall, three studies reporting five models showed low risk of bias, and one study reporting two models for recidivism prediction with different follow-up periods (Wang et al., 2023) carried out all required procedures to minimise the risk of bias. The reported models were based

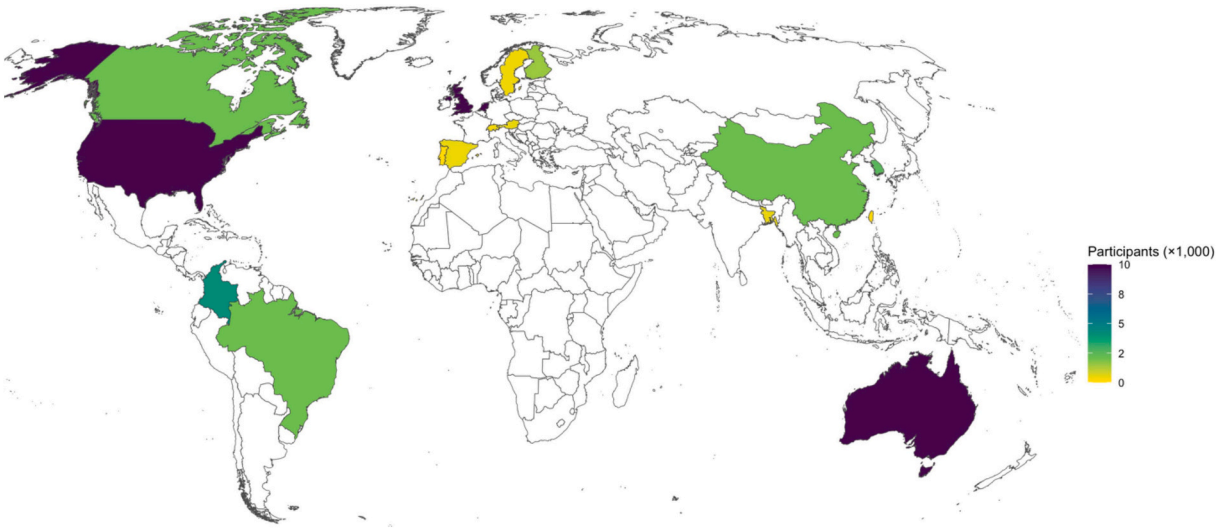

Fig. 4. Geographic coverage of study samples in participant numbers in thousands.

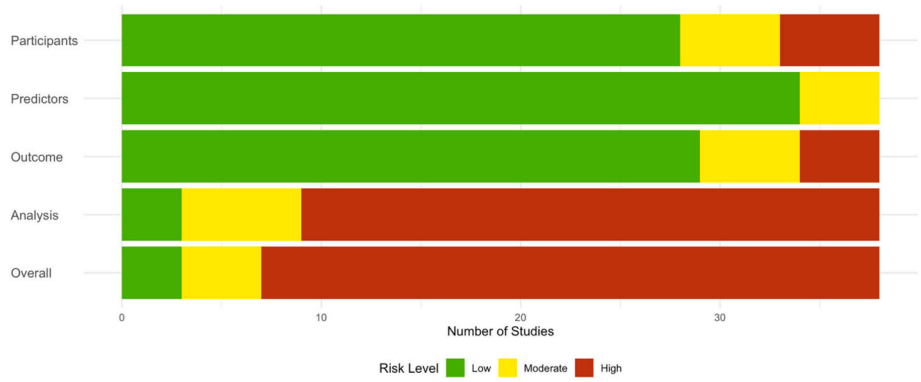

Fig. 5. Risk of bias in the included studies.

on SVM and Gradient Boosting methods, and the studies were primarily led by computer scientists.

4. Discussion

This systematic review summarises over 30 years of research in violence prediction modelling using black box machine learning algorithms. This synthesis provides an overview of 40 models, with most research focused on model development. We have included 38 studies covering over two million participants. There has been substantial research expansion in the field and increasing use of deep learning algorithms, including natural language processing.

In general, published models showed good discrimination performance, but reporting of calibration is scarce, despite its importance in evaluating prediction models (Van Calster et al., 2019). This does not align with current reporting guidelines, such as TRIPOD-AI (Collins et al., 2024). The quality assessment reported in this review also finds other limitations in included studies, particularly in the analysis domain. Reliance on discrimination as the primary evaluation metric, even where calibration performance is additionally reported, does not constitute a methodologically robust assessment. As most included studies demonstrated a high risk of bias, performance findings reported in the primary ML literature should be interpreted with considerable caution. One key methodological shortcoming that had led to overfitting is that datasets were not large enough for sufficient model training. In addition, some models were based on retrospective data and did not clarify the separation between predictors and outcomes, and a data-driven approach was used for variable selection, which may have led to potential biases. External validation was conducted in only three studies, with one (Wang et al., 2023) concluding limited generalisability. Even high-quality models are recommended to be validated for the specific setting in which they are planned to be used (e.g. external geographic validation for COMPAS algorithm in the USA) (Engel et al., 2025). Due to these methodological limitations, high risk of bias, and lack of generalisability and transparency, none of the included models can be recommended for implementation.

Based on these findings, we propose five key considerations for future research on violence prediction modelling. First, a higher level of adherence to quality standards and more interdisciplinary collaboration is required. We recommend that researchers routinely use quality assessment tools to evaluate risk of bias in model development and follow reporting guidelines closely. We also noted large variations in interdisciplinary collaboration across study teams. A minority involved collaborations between clinicians and computer scientists. In studies led mainly by clinicians, analytic methods were at greater risk of bias. We noted some imprecision in terminology (e.g., mixing “classification” and “prediction”) and a limited discussion of assumptions and other key features of complex algorithms. Higher quality studies led by computer scientists frequently did not address clinical problems but rather explored outputs of using a certain method. More collaboration between practitioners (in health and/or criminal justice) and computer scientists will ensure that well-trained models are used to solve high-priority issues in the field. Consideration should also be given to involving ethicists and statisticians with methodological expert in prediction

modelling. Therefore, our first recommendation is for routine assessment of model quality and interdisciplinary collaboration in model development.

Second, most of the data used to develop models were from routine administrative and/or clinical data, with no complex interactions or high-dimensional features. An established principle of clinical prediction models is to use simpler methods (e.g. Cox proportional hazard models, logistic regression) for the structured data with a moderate number of predictors (Steyerberg, 2019). The use of complex black box models for modelling on such datasets does not typically lead to improved model performance compared with the classic statistical methods (Christodoulou et al., 2019), also reported in one of the included studies (Etzler et al., 2024). Moreover, logistic regression models usually calibrate better on routine data (Kantidakis et al., 2023) and provide more stable results (Riley & Collins, 2023). Our second recommendation is to reserve the use of black box models for large, complex multidimensional datasets.

One way to improve the accuracy of model prediction is to use temporal proximity, where possible. Risk of violence is dynamic and is based on different individual-level domains. For example, risk of reoffending decays with time, lowering to a level comparable to the general population (Hanson, 2018). Additionally, studies suggest that various individual-level factors have different periods in which they contribute to risk of violence, and some are associated with the higher risks of violence in the first week after occurrence (Sariaslan et al., 2016), hence risk scores need to be frequently updated to provide the highest accuracy (Lee et al., 2024). Temporal proximity and dynamic predictions may be implemented in mental health with the Cox hazards regression with sliding window landmarks (Houwelingen & Putter, 2011; Yukhnenko et al., 2025), and some studies on dynamic prediction with black box models and its implementation in healthcare are promising (Devaux et al., 2022; Lin & Luo, 2022; Tanner et al., 2021). If the data allow for modelling dynamic predictions, a third research recommendation is to prioritise this approach to test whether it leads to higher precision.

Limited understanding of reaching the decision by models creates a lack of trust and hesitation to implement even high performing algorithms with low risk of bias. In clinical settings, risk estimates are the most beneficial when linked to modifiable factors. In criminal justice, transparency of decision-making is an important feature, and attracted considerable discussion, including how to limit systemic biases and ensure fairness (Farayola et al., 2025). Therefore, a further recommendation is to consider using methods that can provide higher transparency. One of these approaches is the use of explainable AI (XAI) (Longo et al., 2024), where additional steps are implemented to highlight reasons why algorithm reached its decision. This approach has been criticised for overcomplicating implementation (McCoy et al., 2022) or for justifying the unnecessary use of black box algorithms (Rudin, 2019). In addition, such methods do not establish causal links between variables and an outcome (Petch et al., 2022). However, explainable AI can build trust with practitioners, which in turn can make models easier for implementation. Developing models with the use of such approaches as the Transparency and Interpretability Framework For Understandability (TIFU) (Joyce et al., 2023) will ensure more transparent and trustworthy systems.

Another approach to enhance transparency and usefulness is causal machine learning, where the algorithm is built based on prespecified structural causal model (Kaddour et al., 2022). Establishing causal relationships would help identify the modifiable risk factors to inform violence prevention and risk management. This would require using experimental or quasi-experimental methods when developing a prediction model, while current model development has mostly been conducted on observational data. The main benefit of using this method is that it estimates individualised causal effects, which may support more personalised decisions, leading to higher applicability of prediction models. The applicability of this method for violence prediction is not known but it shows promise in other healthcare related predictions (Feuerriegel et al., 2024). Not all decisions predicted by algorithms require transparent causal links: accurate identification may mean that resources for the triage and assessment are allocated effectively (Seyedsalehi et al., 2025). However, in situations requiring causal inference and where data allow, using causal machine learning may be appropriate.

These five recommendations require a comprehensive understanding of both machine learning algorithms and the field of application. Ensuring that all features of the data have been considered, a model has been developed and validated with the adherence to high quality, and a prediction benefits people working in criminal justice and related health and other services, and people involved with the justice system, requires strong interdisciplinary collaboration. The results of this study correspond with findings of previous systematic reviews on violence and crime outcomes. The evidence of mixed performance and methodological shortcomings have been reported for violence risk prediction tools used in forensic mental health settings (Ogonah et al., 2023) and at the sentencing stage in criminal justice (Fazel et al., 2022), regardless of whether advanced ML methods or traditional regression methods were used.

This study provides a comprehensive overview of existing machine learning models developed for violence prediction. However, there are limitations. Systematically searching grey literature was difficult due to lack of databases with these studies, and some reports may have been overlooked, despite using broad terms in our search strategy and conducting additional manual search. Included studies were highly heterogeneous precluding meta-analysis. A broad scope of the review also precluded the in-depth discussion of outcome-specific features.

5. Conclusion

Despite rapid progression in the use of machine learning for violence prediction, methodological shortcomings, limited geographic coverage, and lack of external validations do not allow evaluation of performance outside of training dataset of included models. Even if models are developed with higher quality and more comprehensive reporting, implementation will require examining model feasibility, acceptability, cost-effectiveness, and impact studies. We have outlined five research recommendations: (i) to ensure higher quality of model development and validation and strong collaboration across relevant disciplines (including health, criminal justice, computer/data science, and ethics); (ii) focusing on developing machine learning algorithms with highly complex data while using traditional statistical methods for routinely collected low-dimensional data; (iii) where possible, testing and incorporating temporal proximity and dynamic predictions; (iv) where black box models are necessary, develop more trustworthy algorithms with explainable AI methods; and (v) in predictions requiring clear causal links, apply causal machine learning methods.

Authors contributor

SK completed data screening, data extraction, data curation, data visualisation, and writing (original draft and editing). GS conducted the abstract, title screening, reviewed data extraction, and assisted in writing. DY and SF led on conceptualisation, supervision, and writing (review and editing). SK and GS directly accessed and verified the underlying data reported in the manuscript. All authors had full access to all the data in the study and accept responsibility to submit for publication.

CRediT authorship contribution statement

Stefaniya Kozhevnikova: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Denis Yukhnenko:** Writing – review & editing, Supervision, Methodology,

Data curation. **Giulio Scola:** Writing – review & editing, Formal analysis, Data curation. **Seena Fazel:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcrimjus.2026.102697.

References

Ahmed, A. A., & Okoroafor, N. (2023). An ML-powered risk assessment system for predicting prospective mass shooting. *Computers, 12*(2), 42. https://doi.org/10.3390/computers12020042

Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=uZnSDwAAQBAJ&oi=fnd&pg=PR7&ots=xOwSrtHmwV&sig=Vw4W7_hgW56Sl6SZpGpH4C2ttr0&redir_esc=y#v=onepage&q&f=false.

Austin, P. C., & Steyerberg, E. W. (2017). Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research, 26*(2), 796–808. https://doi.org/10.1177/0962280214558972

Barzman, D., Hemphill, R., Appel, K., Kerekes, O., Sorter, M., Berry, A., … Lin, P. D. (2022). A large naturalistic study on the BRACHA: Confirmation of the predictive validity. *The Psychiatric Quarterly, 93*(3), 803–811. https://doi.org/10.1007/s11126-022-09993-4

Baúto, R. V., Cardoso, J., & Leal, I. (2024). Artificial neural network model for predicting child sexual offending: Role of cognitive distortions, sexual coping, and attitudes. *Journal of Forensic Psychology Research and Practice, 24*(5), 752–770. https://doi.org/10.1080/24732850.2023.2249518

Berk, R., & Bleich, J. (2014). Forecasts of violence to inform sentencing decisions. *Journal of Quantitative Criminology, 30*(1), 79–96. Retrieved from https://www.jstor.org/stable/43551980.

Berk, R. A., & Sorenson, S. B. (2020). Algorithmic approach to forecasting rare violent events. *Criminology & Public Policy, 19*(1), 213–233. https://doi.org/10.1111/1745-9133.12476

Bokhari, E., & Hubert, L. (2018). The lack of cross-validation can lead to inflated results and spurious conclusions: A re-analysis of the MacArthur violence risk assessment study. *Journal of Classification, 35*(1), 147–171. https://doi.org/10.1007/s00357-018-9252-3

Centre for Reviews and Dissemination. (2009). *Systematic reviews: CRD'S guidance for undertaking reviews in health care*. York: CRD.

Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine — Beyond the peak of inflated expectations. *New England Journal of Medicine, 376*(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071

Cheng, N., Guo, M., Yan, F., Guo, Z., Meng, J., Ning, K., … Wang, C. (2023). Application of machine learning in predicting aggressive behaviors from hospitalized patients with schizophrenia. *Frontiers in Psychiatry, 14*. https://doi.org/10.3389/fpsyt.2023.1016586

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, Article e078378. https://doi.org/10.1136/bmj-2023-078378

Devaux, A., Genuer, R., Peres, K., & Proust-Lima, C. (2022). Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history: A landmark approach. *BMC Medical Research Methodology, 22*(1), 188. https://doi.org/10.1186/s12874-022-01660-3

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Paper presented at the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Dobbins, N. J., Chipkin, J., Byrne, T., Ghabra, O., Siar, J., Sauder, M., … Black, T. M. (2024). Deep learning models can predict violence and threats against healthcare providers using clinical notes. *Npj Mental Health Research, 3*(1), 61. https://doi.org/10.1038/s44184-024-00105-7

Engel, C., Linhardt, L., & Schubert, M. (2025). Code is law: How COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law, 33*(2), 383–404. https://doi.org/10.1007/s10506-024-09389-8

Etzler, S., Schönbrodt, F. D., Pargent, F., Eher, R., & Rettenberger, M. (2024). Machine learning and risk assessment: Random forest does not outperform logistic regression in the prediction of sexual recidivism. *Assessment, 31*(2), 460–481. https://doi.org/10.1177/10731911231164624

Fajar, A., Yazid, S., & Budi, I. (2024). *Comparative analysis of black-box and white-box machine learning model in phishing detection*. https://doi.org/10.48550/arXiv.2412.02084

Farayola, M. M., Tal, I., Saber, T., Connolly, R., & Bendechache, M. (2025). A fairness-focused approach to recidivism prediction: Implications for accuracy, trust, and equity. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02452-1

Fazel, S., Burghart, M., Fanshawe, T., Gil, S. D., Monahan, J., & Yu, R. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. *Journal of Criminal Justice, 81*, Article 101902. https://doi.org/10.1016/j.jcrimjus.2022.101902

Fernandez-Felix, B. M., López-Alcalde, J., Roqué, M., Muriel, A., & Zamora, J. (2023). CHARMS and PROBAST at your fingertips: A template for data extraction and risk of bias assessment in systematic reviews of predictive models. *BMC Medical Research Methodology, 23*(1), 44. https://doi.org/10.1186/s12874-023-01849-0

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., … van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. *Nature Medicine, 30*(4), 958–968. https://doi.org/10.1038/s41591-024-02902-1

Garcia-Vergara, E., Almeda, N., Fernández-Navarro, F., & Becerra-Alonso, D. (2023). Artificial intelligence extracts key insights from legal documents to predict intimate partner femicide. *Scientific Reports, 13*(1), 18212. https://doi.org/10.1038/s41598-023-45157-5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.) Retrieved from https://bvbr.bib-bvb.de:443/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=029230557&sequence=000003&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.

Gordon, J. S. (1992). A neural network approach to the prediction of violence (Ph.D.). Available from ProQuest Dissertations & Theses Global. (304013031). Retrieved from https://www.proquest.com/dissertations-theses/neural-network-approach-prediction-violence/docview/304013031/se-2?accountid=13042 https://oxford.primo.exlibrisgroup.com/openurl/44OXF_INST/44OXF_INST:SOLO?genre=dissertations&issn=&title=A+neural+network+approach+to+the+prediction+of+violence&volume=&issue=&date=1992&atitle=&spage=&sid=ProQuest+Dissertations+%2526+Theses+Global&author=Gordon.

Grann, M., & Långström, N. (2007). Actuarial assessment of violence risk: To weigh or not to weigh? *Criminal Justice and Behavior, 34*(1), 22–36. https://doi.org/10.1177/0093854806290250

Hamilton, Z., Neuilly, M., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology, 11*(2), 299–318. https://doi.org/10.1007/s11292-014-9221-8

Hanson, R. K. (2018). Long-term recidivism studies show that desistance is the norm. *Criminal Justice and Behavior, 45*(9), 1340–1346. https://doi.org/10.1177/0093854818793382

Hartman, N., Kim, S., He, K., & Kalbfleisch, J. D. (2023). Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine, 42*(13), 2179–2190. https://doi.org/10.1002/sim.9717

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844. https://doi.org/10.1109/34.709601

Hoessly, L. (2025). On misconceptions about the brier score in binary prediction models. Retrieved from http://arxiv.org/abs/2504.04906.

Hossain, M. M., Asadullah, M., Rahaman, A., Miah, M. S., Hasan, M. Z., Paul, T., & Hossain, M. A. (2021). Prediction on domestic violence in Bangladesh during the COVID-19 outbreak using machine learning methods. *Applied System Innovation, 4*(4), 77. https://doi.org/10.3390/asi4040077

Houwelingen, H. V., & Putter, H. (2011). *Dynamic prediction in clinical survival analysis*. Boca Raton: CRC Press. https://doi.org/10.1201/b11311. Retrieved from https://www.taylorfrancis.com/books/mono/10.1201/b11311/dynamic-prediction-clinical-survival-analysis-hans-van-houwelingen-hein-putter.

Hu, Y., Hung, J., Hu, L., Huang, S., & Shen, C. (2023). An analysis of Chinese nursing electronic medical records to predict violence in psychiatric inpatients using text mining and machine learning techniques. *PLoS One, 18*(6), Article e0286347. https://doi.org/10.1371/journal.pone.0286347

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association, 27*(4), 621–633. https://doi.org/10.1093/jamia/ocz228

Ivaskevics, K., & Haller, J. (2022). Risk matrix for violent radicalization: A machine learning approach. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.745608

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo*. https://doi.org/10.5555/2074158.2074196, 1995.

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine, 6*(1), 6. https://doi.org/10.1038/s41746-023-00751-9

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). *Causal machine learning: A survey and open problems*. Retrieved from http://arxiv.org/abs/2206.15475.

Kantidakis, G., Putter, H., Litière, S., & Fiocco, M. (2023). Statistical models versus machine learning for competing risks: Development and validation of prognostic models. *BMC Medical Research Methodology, 23*(1), 51. https://doi.org/10.1186/s12874-023-01866-z

Karystianis, G., Cabral, R. C., Han, S. C., Poon, J., & Butler, T. (2021). Utilizing text mining, data linkage and deep learning in police and health records to predict future offenses in family and domestic violence. *Frontiers in Digital Health, 3*. https://doi.org/10.3389/fdgth.2021.602683

Kim, K., Yang, Y., Kim, M., Park, J. S., & Kim, J. (2023). Predicting adolescent violence in wartegg-ZeichenTest drawing images based on deep learning. *Connection Science, 35*(1), Article 2286186. https://doi.org/10.1080/09540091.2023.2286186

Laqueur, H. S., & Copus, R. W. (2024). An algorithmic assessment of parole decisions. *Journal of Quantitative Criminology, 40*(1), 151–188. https://doi.org/10.1007/s10940-022-09563-8

Lawn, R. B., & Koenen, K. C. (2022). Homicide is a leading cause of death for pregnant women in US. *BMJ, 379*, Article o2499. https://doi.org/10.1136/bmj.o2499

Le, D. V., Montgomery, J., Kirkby, K. C., & Scanlan, J. (2018). Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *Journal of Biomedical Informatics, 86*, 49–58. https://doi.org/10.1016/j.jbi.2018.08.007

Lee, S. C., Babchishin, K. M., Mularczyk, K. P., & Hanson, R. K. (2024). Dynamic risk scales degrade over time: Evidence for reassessments. *Assessment, 31*(3), 698–714. https://doi.org/10.1177/10731911231177227

Lin, J., & Luo, S. (2022). Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in Medicine, 41*(15), 2894–2907. https://doi.org/10.1002/sim.9392

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology, 27*(4), 547–573. https://doi.org/10.1007/s10940-011-9137-7

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., … Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion, 106*, Article 102301. https://doi.org/10.1016/j.inffus.2024.102301

McCoy, L. G., Brenna, C. T. A., Chen, S. S., Vold, K., & Das, S. (2022). Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology, 142*, 252–257. https://doi.org/10.1016/j.jclinepi.2021.11.001

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry, 27*(6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4

Menger, V., Scheepers, F., & Spruit, M. (2018). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences, 8*(6), 981. https://doi.org/10.3390/app8060981

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill. Retrieved from https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf.

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Navarro, C. A., Dhiman, P., … Smeden, M. V. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, Article e082505. https://doi.org/10.1136/bmj-2024-082505

Moons, K. G. M., Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., … Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. *PLoS Medicine, 11*(10), Article e1001744. https://doi.org/10.1371/journal.pmed.1001744

Neuilly, M., Zgoba, K. M., Tita, G. E., & Lee, S. S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies, 15*(2), 154–176. https://doi.org/10.1177/1088767911406867

Ogonah, M. G. T., Seyedsalehi, A., Whiting, D., & Fazel, S. (2023). Violence risk assessment instruments in forensic psychiatric populations: A systematic review and meta-analysis. *The Lancet Psychiatry, 10*(10), 780–789. https://doi.org/10.1016/S2215-0366(23)00256-0

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Parmigiani, G., Barchielli, B., Casale, S., Mancini, T., & Ferracuti, S. (2022). The impact of machine learning in predicting risk of violence: A systematic review. *Frontiers in Psychiatry, 13*. https://doi.org/10.3389/fpsyt.2022.1015914

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Pelham, W. E., Petras, H., & Pardini, D. A. (2020). Can machine learning improve screening for targeted delinquency prevention programs? *Prevention Science, 21*(2), 158–170. https://doi.org/10.1007/s11121-019-01040-2

Petch, J., Di, S., & Nelson, W. (2022). Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology, 38*(2), 204–213. https://doi.org/10.1016/j.cjca.2021.09.004

Petering, R., Um, M., Alipoufard, N., Tavabi, N., Kumari, R., & Gilani, S. N. (2018). Artificial intelligence to predict intimate partner violence perpetration. In E. Rice, & M. Tambe (Eds.), *Artificial intelligence and social work* (pp. 195–210). Cambridge: Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/artificial-intelligence-and-social-work/artificial-intelligence-to-predict-intimate-partner-violence-perpetration/E16BFC76AFD8BEF09BE8677617E4ADA6.

Ramesh, T., Igoumenou, A., Vazquez Montes, M., & Fazel, S. (2018). Use of risk assessment instruments to predict violence in forensic psychiatric hospitals: A systematic review and meta-analysis. *European Psychiatry, 52*, 47. https://doi.org/10.1016/j.eurpsy.2018.02.007

Riley, R. D., & Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal, 65*(8), Article 2200302. https://doi.org/10.1002/bimj.202200302

Rokach, L., & Maimon, O. (2013). *Data mining with decision trees*. World Scientific. https://doi.org/10.1142/9097. Retrieved from https://doi.org/10.1142/9097.

Rosellini, A. J., Monahan, J., Street, A. E., Heeringa, S. G., Hill, E. D., Petukhova, M., … Kessler, R. C. (2016). Predicting non-familial major physical violent crime perpetration in the US army from administrative data. *Psychological Medicine, 46*(2), 303–316. https://doi.org/10.1017/S0033291715001774

Rosellini, A. J., Stein, M. B., Benedek, D. M., Bliese, P. D., Chiu, W. T., Hwang, I., … Kessler, R. C. (2018). Predeployment predictors of psychiatric disorder-symptoms and interpersonal violence during combat deployment. *Depression and Anxiety, 35*(11), 1073–1080. https://doi.org/10.1002/da.22807

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salo, B., Laaksonen, T., & Santtila, P. (2019). Predictive power of dynamic (vs. static) risk factors in the finnish risk and needs assessment form. *Criminal Justice and Behavior, 46*(7), 939–960. https://doi.org/10.1177/0093854819848793

Santamaría-García, H., Baez, S., Aponte-Canencio, D. M., Pasciarello, G. O., Donnelly-Kehoe, P. A., Maggiotti, G., … Ibáñez, A. (2021). Uncovering social-contextual and individual mental health factors associated with violence via computational inference. *Patterns, 2*(2). https://doi.org/10.1016/j.patter.2020.100176

Santos, R.d. V.d., Coelho, J. V. V., Cacho, N. A. A., & de Araújo, D. S. A. (2024). A criminal macrocause classification model: An enhancement for violent crime analysis considering an unbalanced dataset. *Expert Systems with Applications, 238*, Article 121702. https://doi.org/10.1016/j.eswa.2023.121702

Sariaslan, A., Lichtenstein, P., Larsson, H., & Fazel, S. (2016). Triggers for violent criminality in patients with psychotic disorders. *JAMA Psychiatry, 73*(8), 796–803. https://doi.org/10.1001/jamapsychiatry.2016.1349

Seyedsalehi, A., Scola, G., & Fazel, S. (2025). Precision psychiatry: Thinking beyond simple prediction models – Enhancing causal predictions: Commentary, Seyedsalehi et al. *The British Journal of Psychiatry*, 1–2. https://doi.org/10.1192/bjp.2025.10473

Snell, K. I. E., Levis, B., Damen, J. A. A., Dhiman, P., Debray, T. P. A., Hooft, L., … Riley, R. D. (2023). Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: Checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ, 381*, Article e073538. https://doi.org/10.1136/bmj-2022-073538

Sonnweber, M., Lau, S., & Kirchebner, J. (2021). Violent and non-violent offending in patients with schizophrenia: Exploring influences and differences via machine learning. *Comprehensive Psychiatry, 107*, Article 152238. https://doi.org/10.1016/j.comppsych.2021.152238

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-16399-0

Tanner, K. T., Sharples, L. D., Daniel, R. M., & Keogh, R. H. (2021). Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(1), 3–30. https://doi.org/10.1111/rssa.12611

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(2), 565–584. https://doi.org/10.1111/j.1467-985X.2012.01056.x

United Nations Office on Drugs and Crime. (2023a). *Victims of intentional homicide*. UNODC. Retrieved from https://dataunodc.un.org/dp-intentional-homicide-victims-est.

United Nations Office on Drugs and Crime. (2023b). *Violent offences*. UNODC. Retrieved from https://dataunodc.un.org/dp-crime-violent-offences.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., Topic Group "Evaluating Diagnostic Tests and Prediction Models" of the STRATOS Initiative, … Vickers, A. J. (2019). Calibration: The achilles heel of predictive analytics. *BMC Medicine, 17*(1), 230.

Verrey, J., Ariel, B., Harinam, V., & Dillon, L. (2023). Using machine learning to forecast domestic homicide via police data and super learning. *Scientific Reports, 13*(1), 22932. https://doi.org/10.1038/s41598-023-50274-2

Walsh, C. G., Sharman, K., & Hripcsak, G. (2017). Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *Journal of Biomedical Informatics, 76*, 9–18. https://doi.org/10.1016/j.jbi.2017.10.008

Wang, C., Han, B., Patel, B., & Rudin, C. (2023). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology, 39*(2), 519–581. https://doi.org/10.1007/s10940-022-09545-w

Wang, K. Z., Bani-Fatemi, A., Adanty, C., Harripaul, R., Griffiths, J., Kolla, N., … De Luca, V. (2020). Prediction of physical violence in schizophrenia with machine learning algorithms. *Psychiatry Research, 289*, Article 112960. https://doi.org/10.1016/j.psychres.2020.112960

Watts, D., Mamak, M., Moulden, H., Upfold, C., de Azevedo Cardoso, T., Kapczinski, F., & Chaimowitz, G. (2023). The HARM models: Predicting longitudinal physical aggression in patients with schizophrenia at an individual level. *Journal of Psychiatric Research, 161*, 91–98. https://doi.org/10.1016/j.jpsychires.2023.02.030

Watts, D., Moulden, H., Mamak, M., Upfold, C., Chaimowitz, G., & Kapczinski, F. (2021). Predicting offenses among individuals with psychiatric disorders - A machine learning approach. *Journal of Psychiatric Research, 138*, 146–154. https://doi.org/10.1016/j.jpsychires.2021.03.026

White, N., Parsons, R., Collins, G., & Barnett, A. (2023). Evidence of questionable research practices in clinical prediction models. *BMC Medicine, 21*(1), 339. https://doi.org/10.1186/s12916-023-03048-6

Whiting, D., Lichtenstein, P., & Fazel, S. (2021). Violence and mental disorders: A structured review of associations by individual diagnoses, risk factors, and risk assessment. *The Lancet Psychiatry, 8*(2), 150–161. https://doi.org/10.1016/S2215-0366(20)30262-5

World Health Organization. (2014). *Global status report on violence prevention 2014*. Geneva: WHO.

Wynants, L., Van Smeden, M., Mclernon, D. J., Timmerman, D., Steyerberg, E. W., & Van Calster, B. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine, 17*(1), 192. https://doi.org/10.1186/s12916-019-1425-3

Yukhnenko, D., Blackwood, N., Lichtenstein, P., & Fazel, S. (2025). Dynamic prediction of reoffending in individuals given community sentences: Development and validation of a novel risk monitoring assessment tool (oxMore). *Law and Human Behavior*. https://doi.org/10.1037/lhb0000641

Yukhnenko, D., Farouki, L., & Fazel, S. (2023). Criminal recidivism rates globally: A 6-year systematic review update. *Journal of Criminal Justice, 88*, Article 102115. https://doi.org/10.1016/j.jcrimjus.2023.102115

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180* (3), 689–722. https://doi.org/10.1111/rssa.12227