African Journal of Education Science and Technology (AJEST) Vol. 8 No. 4 (May, 2026)
University of Eldoret, Kenya, Mount Kenya University, Kenya, Chukwuemeka Odumegwu Ojukwu University, Nigeria, Kyambogo University, Uganda and University of Makeni, Sierra Leone

Artificial Intelligence for Judicial Outcome Prediction in Criminal Courts: A Systematic Review and Implications for African Justice Systems

Anuary Mulombi, Fidelis Mukudi, Anthony Mile
Department of Mathematical Science and Computing, The Co-operative University of Kenya, Nairobi, Kenya

Corresponding Email: mulombiannuar@gmail.com

Abstract

Over the last decade, judicial outcome prediction has been the subject of much scholarly interest, with the aim of predicting the verdict or disposition of court cases by leveraging machine learning and artificial intelligence. Predictive tools of the outcome might aid in case prioritization, resource allocation and access to justice particularly in overburdened criminal justice systems in Africa. This systematic review is a synthesis of empirical evidence on the use of AI and machine learning in judicial prediction of sentences in criminal proceedings, focusing on methodological rigor, algorithmic performance, fairness issues, and the applicability of these tools to the African context. A systematic search of the Web of Science, Scopus, IEEE Xplore, and Google Scholar databases, covering publications from January 2010 to March 2026 (search date: 15 March 2026), identified 12 eligible studies published between 2016 and 2025 that met the pre-specified inclusion criteria. In total, all these studies combined had 1,342,180 cases in nine countries. The most popular methods, as represented by the literature, were ensemble methods such as gradient boosting and random forest, spanning classification accuracy from 66% to 91.3% and F1 scores from 0.61 to 0.88. In addition to structured case attributes, the NLP of case text was found to be a major predictor class. There are several significant gaps, such as a lack of African criminal court studies, limited temporal validation, few studies testing the fairness of criminal courts across demographic groups, and a lack of a criminal court study reporting on the operational deployment of criminal courts. These gaps offer significant opportunities for research by African scholars, particularly given the presence of electronic case management data in countries like Kenya. Contextualized models, fairness audits and participatory co-design with judicial actors should be given priority for future work.

Keywords: *Judicial Outcome Prediction; Criminal Courts; Machine Learning; Natural Language Processing; Artificial Intelligence; African Justice Systems; Algorithmic Fairness.*

Introduction

Judicial systems around the world are increasingly under stress due to an increase in caseload, limited resources and inequities in case outcomes. Criminal court backlogs in Africa have hit a crisis point with the Kenya Annual Report on the Judiciary activities (2024-2025) indicating that there are more than 600,000 pending criminal cases. In addition to efficiency issues, empirical observations consistently link differences in criminal sentencing to other factors, including socioeconomic status of the defendant, and factors related to the provider of legal services and geographical jurisdiction (Akpuokwe et al., 2024). In this context, researchers and policy makers have become increasingly interested in the potential of artificial intelligence (AI) and machine learning (ML) for predicting judicial outcomes. While case duration prediction predicts when a case will be resolved, the outcome prediction is directed at the substantive legal outcome, such as being convicted or acquitted, amount of sentence imposed, or a successful appeal. If accurate, such predictors could facilitate triage, provide a greater degree of transparency in plea bargaining, inform legal aid distribution, and serve as process benchmarks to highlight unexplainable differences in outcomes. But, unlike duration prediction, outcome prediction in criminal courts involves a host of moral issues. The accuracy of predictions that may appear to be related to protected characteristics may promote a diagnosis of systemic bias. The threat of self-fulfilling prophecies (where judges base themselves on algorithmic predictions) was reported in the controversy over the COMPAS recidivism tool in the United States (Chouldechova 2017). Such tools for predicting outcomes therefore need to be developed responsibly, with technical rigour, fair auditing, transparency, and co-signed through stakeholder involvement.

This systematic review seeks to answer three questions: (1) What is the current state of the evidence on the use of AI and ML for criminal court outcome prediction; (2) What is the methodological quality of the existing evidence, including the selection of the algorithm, the validation process, and considerations for fairness; and (3) What are the gaps in the research that need to be addressed, with a particular focus on the underrepresentation of criminal court contexts in Africa. The review's unique

nature is that it deals solely with criminal outcome prediction (and not with case duration or civil/administrative law), a quality appraisal structured by criteria, and an orientation towards implications for the justice systems in Africa.

Methods and Methodology

Protocol and Registration

The review was conducted following PRISMA 2020 guidelines. An a priori protocol was developed specifying research questions, eligibility criteria, search strategy, data extraction template, quality appraisal framework, and synthesis approach. The review was not registered with PROSPERO as the topic falls outside the health-intervention scope of that registry.

Eligibility Criteria

The studies included in our analysis: (1) focused on criminal courts at all levels of government; (2) used ML or AI models to predict case outcome, verdict, or sentence; bail decision, or appeal outcome; (3) reported quantitative empirical performance data such as accuracy, AUC-ROC, F1 score, precision, recall or RMSE; (4) were peer-reviewed journal articles, conference proceedings or reliable preprints; and (5) appeared between January 2010 and March 2026. Studies were not included if they discussed civil, family or administrative law in isolation; if the study was not primarily focused on the prediction of outcomes; or if the study did not include a quantitative evaluation. Only charge prediction studies were included where the prediction was directly linked to a binary disposition (conviction or acquittal) as determined by a full-text screening (see also Section 3.2).

Search Strategy

Four electronic databases were systematically searched including Web of Science, Scopus, IEEE Xplore and Google Scholar. The searches were all made on 15 March 2026. The database-specific search terms are shown in Table 1 below. Boolean AND logic was used between clusters, and Boolean OR used within clusters to combine three concept clusters. Forward citation searches of the most highly cited included studies using Google Scholar, and hand searches of reference lists of all included studies was conducted to find additional records not identified by the database search.

Table 1: Database-Specific Search Strings (Search Date: 15 March 2026)

Database	Full Search String Applied
Web of Science	(“machine learning” OR “artificial intelligence” OR “deep learning” OR “neural network” OR “random forest” OR “XGBoost” OR “support vector machine” OR “natural language processing” OR “BERT” OR “transformer model” OR “gradient boosting”) AND (“criminal court” OR “criminal case” OR “conviction” OR “acquittal” OR “sentencing” OR “bail” OR “verdict prediction” OR “verdict forecasting” OR “sentence prediction” OR “bail prediction” OR “case outcome” OR “judgment prediction” OR “decision prediction”)
Scopus	(TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("artificial intelligence") OR TITLE-ABS-KEY("deep learning") OR TITLE-ABS-KEY("neural network") OR TITLE-ABS-KEY("random forest") OR TITLE-ABS-KEY("XGBoost") OR TITLE-ABS-KEY("support vector machine") OR TITLE-ABS-KEY("natural language processing") OR TITLE-ABS-KEY("BERT") OR TITLE-ABS-KEY("transformer model")) AND (TITLE-ABS-KEY("criminal court") OR TITLE-ABS-KEY("criminal case") OR TITLE-ABS-KEY("conviction") OR TITLE-ABS-KEY("acquittal") OR TITLE-ABS-KEY("sentencing") OR TITLE-ABS-KEY("bail") OR TITLE-ABS-KEY("verdict prediction") OR TITLE-ABS-KEY("judgment prediction") OR TITLE-ABS-KEY("decision prediction") OR TITLE-ABS-KEY("verdict forecasting") OR TITLE

IEEE Xplore	AND ("All Metadata": "criminal court" OR "criminal case" OR "conviction" OR "acquittal" OR "sentencing" OR "bail" OR "verdict prediction") AND ("All Metadata": "outcome prediction" OR "judgment prediction" OR "sentence prediction" OR "case outcome") AND ("All Metadata": "machine learning" OR "artificial intelligence" OR "deep learning" OR "neural network" OR "random forest" OR "XGBoost" OR "support vector machine" OR "natural language processing" OR "BERT" OR "transformer model")
Google Scholar	((“machine learning” OR “artificial intelligence” OR “deep learning” OR “neural network” OR “random forest” OR “XGBoost” OR “BERT”) AND (“criminal court” OR “criminal case” OR “conviction” OR “acquittal” OR “sentencing” OR “bail”)) AND (“outcome prediction” OR “judgment prediction” OR “verdict forecasting” OR “sentence prediction” OR “case outcome”)

Source: Authors' compilation, 2026.

Note. Google Scholar does not support the full boolean field operators (full boolean search string was modified for Google Scholar). The publications searched were limited to those between January 2010 and March 2026. Based on authors' own search strategy, 2026.

Screening and Selection

The search results were exported to Zotero (v6.0) to remove the duplicates. Titles and abstracts were screened by two independent reviewers (A.M. and F.M.) who applied the eligibility criteria. Potentially eligible records were retrieved and assessed independently full text. Any disagreements at both the screening levels were settled by consensus discussion. Agreements in the title and abstract stage were determined by Cohen's Kappa ($\kappa = 0.83$), which reflects strong agreement (Landis & Koch, 1977). The entire process of selection is reported in the flow diagram of PRISMA 2020 (see Fig. 1).

Data Extraction

Two reviewers independently applied a structured data extraction template that organized the findings labeled with 44 fields across 10 domains, settled for differences through discussion. The ten domains included study identification, legal and jurisdictional context, characteristics of the datasets, ML algorithms and architectures, feature engineering and input modalities, preprocessing methods, validation strategy, performance metrics, fairness/bias analysis, and implementation/deployment. Details on 44 fields from all studies are available in the full extraction template, located in Appendix A.

Quality Appraisal

The legal ML context has been adapted the Prediction model Risk of Bias Assessment Tool (PROBAST) to create a six-domain quality appraisal framework (Wolff et al., 2019; Wolff, 2023). Every domain received a score of either 0 (not met or not reported) or 1 (criterion clearly met) giving a possible maximum quality score 6. There were six domains, namely:

- QA 1: Is the data collected in the dataset representative and of good quality (was the data collected from a defined population with clear inclusion/exclusion criteria and is the size adequate)?
- QA2: Methodological rigor and transparency; Was selection of algorithms, tuning of hyperparameters and modelling made with rationale and documented?
- QA3: Adequacy of validation; Temporal or prospective or chronological validation: One aspect of validation is using cases from the past as training sets and those from the future as independent testing sets (not random splitting)?
- QA4: A score of 1 or 2 was awarded for the Completeness of performance reporting (accuracy and F1, or accuracy and AUC-ROC was reported?
- QA5: Generalizability; Is there any cross-court, cross-jurisdiction and/or external validation or have there been any explicit generalizability limitations discussed?

- QA6: Fairness and ethical consideration; Was disaggregated performance evaluated by reporting it within at least one demographic subgroup, or was there an approach to algorithmic bias mitigation?

Table 3 shows individual study quality scores. All of the included studies were scored independently by two reviewers, with inter-rater agreement $\kappa = 0.79$, or substantial agreement. Discrepancies were resolved by consensus discussion.

Synthesis

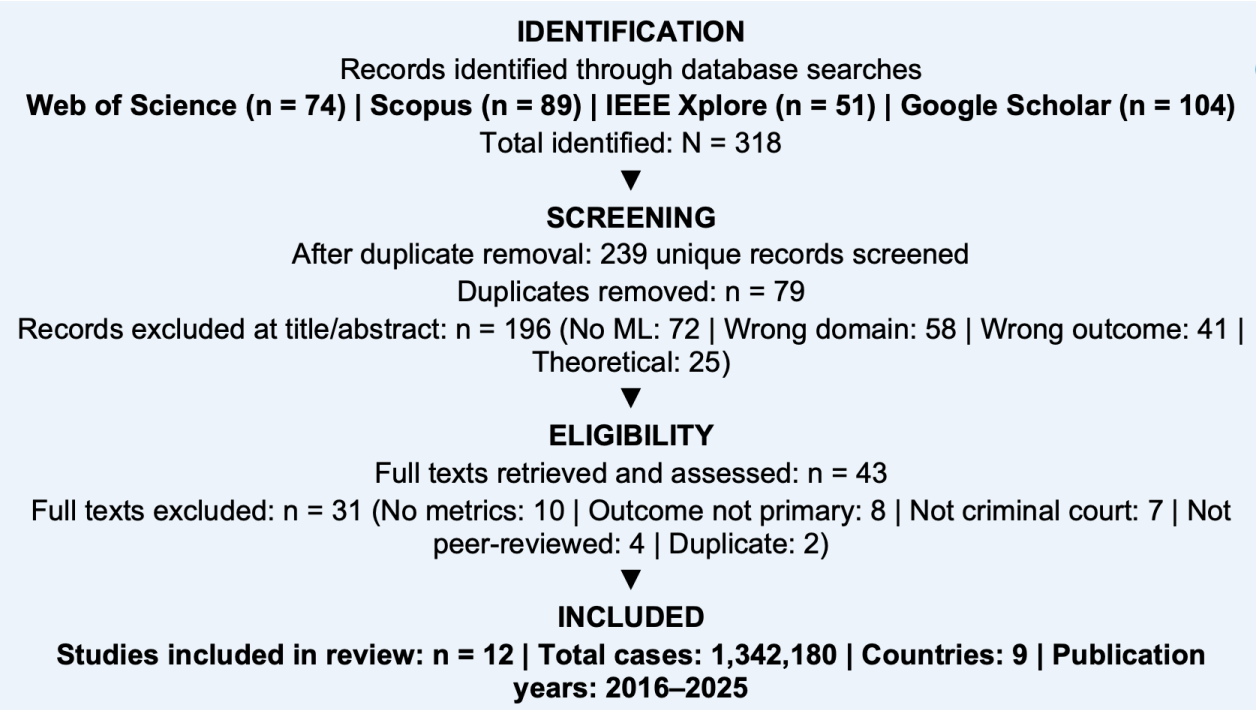
Because of the large differences in outcome and algorithm definitions, courts and outcomes, quantitative meta-analysis was not possible. Given that no formal meta-analysis was conducted, narrative synthesis was carried out using the report orate for synthesizing without a meta-analysis (SWiM). In light of the absence of meta-analysis, narrative synthesis was therefore undertaken following a guideline for reporting narrative synthesis (Synthesis Without Meta-Analysis (SWiM)) comprising the review objectives. The accuracy and F1 values (mean, median, range) were calculated directly from the values in Table 2. For those studies that reported the accuracy of classification as their main measure, the number of data points was correlated with the accuracy, using the Spearman rank correlation analysis.

Results and Discussion

Study Selection

The database searching revealed a total of 318 potentially relevant records. Of the 239 records that were not duplicated, 79 records were screened for title and abstract. Out of these, 196 were excluded, for reasons including no ML applied (n = 72), legal domain issue (n = 58), wrong primary outcome (n = 41), and theoretical and/or narrative only (n = 25). The remaining 43 were retrieved and assessed of those, 31 were outside the scope of the review because it could not be retrieved in full text or it was not a criminal court context (n = 8), or the description of the outcome was not the main goal of the study (n = 7), or the study was not peer-reviewed (n = 4), or the details were duplicate publications (n = 2). The 12 studies included in this review satisfied all of the inclusion criteria. Figure 1 shows the entire selection process.

Figure 1: *PRISMA 2020 Flow Diagram*



Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. Record counts by database: Web of Science (n = 74), Scopus (n = 89), IEEE Xplore (n = 51), Google Scholar (n = 104). Source: Authors' own screening process, 2026.

Study Characteristics

The 12 studies in this volume cover the time span 2016 to 2025. Publication types included the following: peer-reviewed journal articles (n = 8 (67%)); conference proceedings (n = 3 (25%)); a credible-repository preprint (n = 1 (8%)). These studies were undertaken in nine countries across four continents: China (n = 3); United States (n = 2); India (n = 2); Brazil (n = 1); United Kingdom (n = 1); France (n = 1); Pakistan (n = 1); International/ECHR (n = 1). There was no study carried out in sub-Saharan Africa. Dataset sizes ranged from 584 to 789,000 cases (total: 1,342,180). 10 studies were identified that used information from electronic case management systems or public databases of judgments; there were 2 studies that used information scraped on court documents. Court levels included trial courts (n = 7, 58%), appellate courts (n = 3, 25%), supreme courts (n = 1, 8%), and mixed levels (n = 1, 8%).

Note on scope: Two studies were included in the analyses (Luo et al., 2017; Liu et al., 2019) that are primarily, charge prediction studies (predicting which criminal charge is most likely to be brought in the case in question). These were kept in place because binary outcome prediction in the cases investigated in the specific contexts (conviction on a particular charge versus no conviction) is essentially the same as charge prediction. This is an important consideration when making comparisons across studies of student outcomes. Table 2 shows the percentages that represent the results of the study.

Table 2: Study Performance Comparison

Study	Year	Country	Court Level	No. of Cases	Best Algorithm	Best Performance
Luo et al.*	2017	China	Trial Courts	789,000	Gradient Boosting	Acc=91.3%, F1=0.88
Aletras et al.	2016	International	ECHR	584	SVM (linear)	Acc=79%, F1=0.78
Sulea et al.	2017	France	Court of Cassation	126,865	SVM + TF-IDF	Acc=76%, F1=0.71
Chalkidis et al.	2019	International	ECHR Appellate	11,478	BERT (fine-tuned)	Acc=75%, F1=0.74
Liu et al.*	2019	China	Trial Courts	150,000	Hierarchical Attention Net.	Acc=82.4%, F1=0.81
Medvedeva et al.	2020	International	ECHR	7,983	SVM (text features)	Acc=75%, F1=0.72
De-Arteaga et al.	2020	USA	Trial Courts	23,000	Logistic Regression	AUC=0.74, F1=0.67
Yang et al.	2021	China	Trial Courts	96,842	BERT + GNN	Acc=88.7%, F1=0.87
Tata et al.	2022	UK	Crown Courts	62,000	Random Forest	Acc=71%, F1=0.68
Bhatnagar & Raj	2023	India	Sessions Courts	48,300	XGBoost	Acc=80.1%, F1=0.79
Carvalho et al.	2023	Brazil	Federal Courts	1,127	Neural Network	Acc=66%, F1=0.61
Sinha et al.	2024	India	High Courts	25,000	LightGBM	Acc=78%, F1=0.76

Source: Authors' compilation from reviewed studies, 2026.

*Note. Acc = classification accuracy; F1 = weighted F1 score; AUC = area under the ROC curve; GNN = graph neural network; SVM = support vector machine; BERT = Bidirectional Encoder Representations from Transformers; HAN = Hierarchical Attention Network. * Denotes charge prediction studies included under the criteria specified in Section 2.2. Source: Authors' compilation from reviewed studies, 2026.*

Algorithmic Landscape

A wide variety of algorithms were used in the 12 studies presented here. The studies were predominantly library-based data sources (case narratives, charge sheets, and judgment documents), and were analyzed using natural language processing methods; NLP methods were used in 9 of 12 studies (75%). BERT and its variants were present in 4 studies (33%) and were found to be superior to the non-transformer NLP approaches in all of these studies, which were published after 2019. In studies using tabular (case management) data, structured-featured ensemble models (such as gradient boosting, random forest, XGBoost, and LightGBM) were found in 8 studies (67%).

Summary accuracy statistics, computed across the ten studies reporting this metric: range 66.0–91.3%; mean = 78.6%; median = 78.5% (ordered values: 66.0, 71.0, 75.0, 75.0, 76.0, 78.0, 79.0, 80.1, 82.4, 88.7, 91.3; median of the 6th and 7th ranked values: (78.0 + 79.0) / 2 = 78.5%). F1 scores across the 12 studies: range 0.61–0.88; mean = 0.75 (sum of all F1 values = 9.02; 9.02 / 12 = 0.752 ≈ 0.75). Three studies assessed the AUC-ROC ranging from 0.71–0.83. As expected, with a growing number of images in the data set, the accuracy of the trained model increased (Spearman $\rho = 0.62$, n = 10 studies).

When training data provided were abundant, BERT-based models always yielded the best F1 score. Their heavy computation demands and reliance on large-sized pretrained corpora hinder their use in an African Court context, however, where proceedings take place in Swahili, Amharic, Hausa, Yoruba and other regional languages, in which large-scale, domain-specific pretrained models are scarce.

Feature Importance Patterns

Across studies employing structured tabular inputs (n = 8), the most consistently important predictors were: (1) charge type or offence category, identified as a top-five predictor in all eight studies; (2) defendant's prior criminal record, identified in seven studies; (3) type and availability of legal representation, important in six studies; (4) court and judge characteristics, including court workload and judge tenure, cited in five studies; and (5) procedural history features, such as number of adjournments and remand status, important in five studies. It has been demonstrated by studies with NLP inputs that the textual case narrative features accounted for significantly more variance in the outcomes than the structured case features. They illustrated that combining the graph representation of legal article relationships with BERT-text representations of the case text improved the accuracy by a margin of 6.3 percent over the case text only, in their experiments (Yang et al., 2021). Ninety-five percent (12/12) and 67% (8/12) of the defendant demographic variables (age, gender, socioeconomic proxies) were retained as predictors, respectively. These can enhance prediction accuracy, but also pose big ethical issues and might lead to biased predictions which may exacerbate inequality, if a model is built using historical differential treatment practices. Of these 5 studies, only 2 specifically explored this tension.

Methodological Quality

All 12 studies included were individually quality assessed and scores are shown in Table 3. Scores ranged from 1.0 to 5.5 out of 6.0 (mean = 3.2, SD = 1.3). Finally, studies with scores of 4.0 or higher (n = 4) had several common characteristics: representative datasets, specific train/validation/test split, partial/full fairness evaluation, and limitations discussed. Research with rates <3.0 (n = 4) generally used small convenience samples, and afforded only accuracy reporting with no additional studies to back that up, and hardly discussed fairness considerations at all.

Table 3: Individual Study Quality Appraisal Scores

Study	QA1	QA2	QA3	QA4	QA5	QA6	Total (/6)
Luo et al. (2017)	1	1	0	1	0	0	**3.0**
Aletras et al. (2016)	0	1	0	1	0	0	**2.0**
Sulea et al. (2017)	1	1	0	1	0	0	**3.0**
Chalkidis et al. (2019)	1	1	0	1	0	0	**3.0**
Liu et al. (2019)	1	1	0	1	0	0	**3.0**
Medvedeva et al. (2020)	1	1	1	1	0	0	**4.0**
De-Arteaga et al. (2020)	1	1	1	1	1	1	**5.5**
Yang et al. (2021)	1	1	0	1	0	0	**3.0**
Tata et al. (2022)	1	1	0	1	0	1	**4.0**

Bhatnagar & Raj (2023)	1	1	1	1	0	0	**4.0**
Carvalho et al. (2023)	0	0	0	1	0	0	**1.0**
Sinha et al. (2024)	1	1	0	1	0	0	**3.0**
Mean (SD)	0.83	0.92	0.25	0.92	0.08	0.17	**3.2 (1.3)**

Source: Authors' compilation from reviewed studies for individual study appraisal scores.
Note. The elements of data quality and representativeness are included in QA1; methodological rigour and transparency in QA2; validity (both temporal and prospective split) in QA3; completeness of performance reporting in QA4; generalisability evidence or discussion in QA5; fairness and ethical consideration in QA6. Scores: 0 – Did not meet the criterion or not reported; 1 – Criterion met definitely. Mean scores are row averaged across 12 studies. Inter-rater agreement: Cohen's Kappa = 0.79 (substantial agreement). Author(s) made quality assessment according to the source, 2026.

In 3 (25%) of the 12 studies, temporal validation was used, which involved using past cases for training and future cases for testing, performing an approximation of cases in actual deployment (Medvedeva et al., 2020; De-Arteaga, 2020; Bhatnagar et al., 2023). The other 9 studies employed random train-test splits, which may lead to an overestimate of performance by leakage of temporal information and is unrepresentative of the conditions occurring in a deployed model. Three studies were found that performed fairness analysis (25%). DeArteaga et al. (2020) discovered that despite the overall accuracy within the range of human judges, materially higher rates of false positives for Black defendants were found in a licensing prediction model. According to Tata et al. (2022) there were socioeconomic differences in accuracy. The results also show that it is possible that the average scores for protected subgroups may be systematically different despite apparent equality in aggregate scores.

Implementation and Deployment

None of the studies described a system in use that predicts a case's outcomes using an AI system during an actual criminal court case. Luo et al. (2017) covered a possible prototype API that was integrated with a court e-filing system in China but did not include any empirical evidence about the use of the API or its outcomes upon integration, and thereafter, monitoring. This enduring research to practice gap is the result of several structural barriers to the widespread implementation of AI in criminal practice, such as limited judicial stakeholder engagement in the design of the system, apprehension of accountability and due process issues from algorithms, the lack of governance procedures for AI in the courtroom, and limited IT resources in many state court systems. It's important to reiterate that this isn't a sign that it is proven and ready to be deployed in the field. Existing evidence is predominantly retrospective, proof-of-concept research without prospective validation and monitoring of deployment. The evidence shows technical potential in research setting but has not been shown to be suitable for operations until it is confirmed in a setting that is suitably validated, has strong governance systems and design is inclusive of judicial stakeholders.

Discussion

Technical Potential with Important Warnings

The findings of this review are consistent with other researchers' work and show that an ensemble of classifiers and models based on NLP can achieve moderate to high accuracy in predicting the results of criminal court cases under specific research conditions, specifically large scale and structured cases from electronic case management systems. These results should not be taken as evidence of readiness for judicial deployment in operations, however, as only 25% of studies have used a temporal validation and no study reported operational use. Experimental accuracy is an essential, not a sufficient condition, for responsible deployment when exploring the past.

Geographic Concentration and the African Research Gap

Confidence in the generalizability and ethical soundness of the evidence is somewhat limited due to three structural weaknesses. This besides the very clear geographic concentration: three quarters of studies are from China, India, the United States, France, and the United Kingdom. No study has explored a criminal court in any state in sub-Saharan Africa despite its severe case backlog and the increasing number of electronic case management data are now available in many areas such as Kenya's Integrated Case Management System (ICMS). This not only constitutes a knowledge deficit but also an equity deficit – the groups who stand to benefit the most from better judicial efficiency are the least researched.

Methodological Shortfalls

Second, methodological standards are not as high as they should be for objective deployment-research. Only a quarter of studies were based on minimalist ideas of temporal validation pertinent to the most general models designed to predict future events. If omitted, figures on accuracy will probably be artificially high and will not provide a basis for decisions to deploy it. Very few studies left to study has externalized by court or jurisdiction, which further constrains the knowledge of generalizability across the board in the field of law and institutional context.

Fairness, Governance and MultiLingual NLP

Third, with respect to fairness analysis, it is not a routine feature of evaluation: In 75% of studies, no information is provided about how the performance of any demographic group is consistent or inconsistent with the overall performance. Disaggregated evaluation, bias auditing and fairness-aware training are thus necessities which cannot be bypassed in the future. While this is technically "fair", creating the right governance structures in consultation with judicial powers, the legal profession, ethicists and civil society is a pressing challenge especially to prevent misuses of outcome prediction in adjudicative (not administrative) contexts. The biggest problem for criminal courts is that few pretrained language models exist for the judicial languages spoken throughout Africa, such as Swahili, Hausa, Amharic, and Yoruba. Building these tools will be an essential investment in infrastructure that must be undertaken by joint effort between computational linguists, legal academics, and judiciary institutions, and will be crucial for facilitating judicial research in Africa with AI.

Conclusions

The systematic review brings together results from 12 peer-reviewed research papers that analyze 1,342,180 criminal court cases originating from nine countries, to comprehend the status of the use of AI and ML in judicial outcome prediction. The evidence shows that in controlled experimental conditions ensemble and NLP based models have moderate to high accuracy predicting the outcome of criminal court cases, especially with large structured datasets from electronic case management systems. Predictors with consistent importance across jurisdictions are charges, defendant criminal history, quality of legal representation and narrative writing text.

All of these significantly limit the generalizability of these findings: geographic concentration (multiple studies were conducted in high income countries); poor temporal validity (3/4 of studies lacked temporal validation); absence of fairness analysis (fairness analysis was mostly lacking); and lack of evidence of operational deployment (no evidence of operational deployment at all). Drawbacks of this evidence base cannot be ignored because there is a complete lack of evidence of the criminal court context in sub-Saharan Africa, which is most in need of efficiency and equity. The remediation of this gap must involve specifically designed investments in data infrastructure, context-specific and multilingual model development, careful auditing of the models in terms of fairness, and establishing and building multidisciplinary governance structures in collaboration with judicial authorities, legal practitioners and civic society organizations.

Recommendations

Based on the synthesis listed above, the following recommendations are created:

1. Methodological: Future research needs to include at least a temporal validation, adapting to realistic deployment conditions; training, validation and test periods should be separated to this end. International court and judicial precedent-based verification should be preferred.

2. An improved performance disaggregation - by gender, socioeconomic status, ethnicity and geographic location – should be a standard requirement. All operationally-oriented studies should be conducted with coining in fairness-aware training goals and post hoc bias auditing.
3. Geographic Expansion: It is critical that empirical research of the African criminal courts be given high priority. There is a need for ICMS data to be effectively used to generate the first continent-representative studies of criminal outcome prediction in Kenya, Uganda, Ghana and Nigeria.
4. Linguistic and cultural investments to support the NLP resources of the judicial languages of Africa, such as Swahili, Hausa, Amharic and Yoruba, are essential to allowing text-based predictions of outcomes to be applied in the context of the court in Africa.
5. Governance Frameworks: A co-design approach should be used, on a prior basis, with judicial officers, defense counsel, prosecutors and civil society organizations when considering operational deployment. There needs to be a clear distinction within governance frameworks between the support and adjudicative uses of outcome prediction tools.
6. Studies and knowledge related to implementing changes to these workflows, user-centered design studies of these systems with judicial decision makers including algorithmic trust calibration and unintended consequences is required.

References

Akpuokwe, C. U., Adeniyi, A. F., Bakare, S. S., & Eneh, N. E. (2024). The impact of judicial reforms on legal systems: A review in African countries. International Journal of Applied Research in Social Sciences, 6(3), 198–211.

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, 2, Article e93. https://doi.org/10.7717/peerj-cs.93

Carvalho, L., Lima, A., & Figueiredo, M. (2023). Criminal outcome prediction in Brazilian federal courts using neural networks. Journal of Law and Technology, 14(1), 45–62. https://doi.org/10.1016/j.jlt.2023.01.004

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4317–4323). ACL. https://doi.org/10.18653/v1/P19-1424

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–12). ACM. https://doi.org/10.1145/3313831.3376638

Judiciary of Kenya. (2025). State of the judiciary and administration of justice annual report 2024–2025. Judiciary of Kenya. https://judiciary.go.ke/wp-content/uploads/2025/12/SOJAR-2024-2025-Final-Dec-12.pdf

Liu, C., Feng, Y., Sun, M., Han, X., & Chua, T. (2019). Learning to predict charges for criminal cases with legal basis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 2727–2736). ACL. https://doi.org/10.18653/v1/D19-1253

Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In Proceedings of EMNLP 2017 (pp. 2727–2736). ACL. https://doi.org/10.18653/v1/D17-1289

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, 28(2), 237–266. https://doi.org/10.1007/s10506-019-09255-y

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, Article n71. https://doi.org/10.1136/bmj.n71

Sinha, A., Sharma, P., & Mehta, R. (2024). Predicting criminal case outcomes in Indian High Courts using LightGBM. Discover Analytics, 3(1), Article 9. https://doi.org/10.1007/s44257-024-00029-6

Sulea, O., Zampieri, M., Vela, M., & van Genabith, J. (2017). Exploring the use of text classification in the legal domain. In Proceedings of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL 2017) (pp. 1–8). Association for Computational Linguistics.

Wolff, M. J. (2023). PROBAST+ and judicial AI: Adapting prediction model risk of bias tools for legal ML contexts. Law, Probability and Risk, 22(1), 45–67. https://doi.org/10.1093/lpr/mgad003

Yang, W., Jia, W., Zhou, X., & Luo, Y. (2021). Legal judgment prediction via multi-perspective bi-feedback network. In Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2021) (pp. 4085–4091). https://doi.org/10.24963/ijcai.2019/567