

Explainable Artificial Intelligence-Based Multimodal Risk Assessment Framework for Predicting Offending Behaviour and Rehabilitation Outcomes Among Individuals with Intellectual Disabilities

*Preethi M¹, Sugantha Mallika S.S², Selva Priya A³ and Dr. Suganya B⁴

¹Assistant Professor (Sr. Gr), Department of Information Technology, Sri Ramakrishna Engineering College (Affiliated to Anna University Chennai), Coimbatore, Tamil Nadu, India. E-mail: preethi.muthukumar@srec.ac.in, Orcid: <https://orcid.org/0000-0003-1106-8863/>

²Assistant Professor (Sl.Gr), Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India. E-mail: sugantha.samuel@srec.ac.in, Orcid: <https://orcid.org/0000-0002-2571-4232>

³Assistant professor (OG), Department of Information Technology, Sri Ramakrishna Engineering College (Affiliated to Anna University Chennai) Coimbatore, Tamil Nadu, India. E-mail: selvapriya7872@gmail.com, Orcid: <https://orcid.org/0009-0000-3439-1562>

⁴Assistant Professor, Department of Computer science and Engineering, School of computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India. E-mail: drsuganyab@veltech.edu.in, Orcid: <https://orcid.org/0000-0002-8143-6898>

Abstract

People with Intellectual Disabilities (ID) are over-represented in forensic and criminal justice settings, but remain under-served by existing static and actuarial measures of risk that are geared toward more general offender populations. These tools do not capture the dynamic, contextual and cognitive nature of the former population, while newer machine learning techniques are criticized for functioning as a black box. We develop a multimodal risk assessment framework based on Explainable Artificial Intelligence, which takes into account heterogeneous data streams - structured clinical and criminological data, behavioral observation time series, neuropsychological and psychometric assessment, and narrative text - to simultaneously predict offending behavior and treatment success in offenders with ID. Our framework consists of modality-specific encoders, adaptive fusion layers that are capable of taking advantage of complementary information, and a temporal self-attention mechanism to model longitudinal trajectories. We bake interpretability into our framework from the very beginning as opposed to treating it as an afterthought and leverage global additive attributions, local Shapley values, and counterfactuals to generate an intelligible explanation of each prediction. This paper also takes into account the issues of class imbalance and fairness of algorithms because recidivism or re-offending cases are less frequent compared to non-recidivism or non-re-offending cases, and there is also an uneven distribution of impact due to the vulnerable nature of the group involved. The paper achieves this by providing the scientific basis for multimodal fusion through the integration of previous empirical research findings, providing a replicable process of evaluation, and identifying the governance requirements needed for deployment of the model.

Keywords Explainable Artificial Intelligence, Multimodal Learning, Intellectual Disability, Risk Assessment, Offending Behaviour, Recidivism, Rehabilitation Outcomes, Data Fusion, Shap, Fairness.

Article History:

Received: 19.06.2026;

Revised: 07.07.2026;

Accepted: 26.08.2026;

Published: 30.09.2026

Introduction

The combination of intellectual disability and the criminal justice system is an area where the risks posed by predictive modelling are exceptionally high and acceptance of opacity is correspondingly low. Individuals with intellectual disabilities are disproportionately present in forensic settings and have an elevated predisposition to negative consequences, including harm to themselves, extended incarceration, and repeat contacts with the law enforcement [14]. In addition, the tools most often used to evaluate their risk of recidivism were developed with general, neurotypical population of offenders in mind, and do not take into account cognitive, communicative, and adaptive function-related differences that affect both criminal behavior and its correction [2].

Such disparity is associated with some practical problems. For example, a scoping review of forensic models of care for individuals with ID showed that, although scheduled risk assessment was applied in 63.5% of the reviewed models, less than 10% of those models involved instruments adapted to neurodevelopmental features [2]. According to the results of the review, 17.3% of the models incorporated outcome evaluation. The limitations of traditional actuarial tools are also compounded by their dependence on fixed factors, like age and criminal background, which reflect past information but do not account for changing factors such as emotional state, stress levels, social interaction, and environment that lead to dynamic risk [20].

This clinical dilemma mirrors an ongoing revolution in criminal justice prediction. In the past few years, machine learning methods have consistently proven to be at least as accurate as more conventional actuarial approaches [5; Travaini et al., 2022], and an increasing number of scholars have started to use ML models in conjunction with explainability methods (Chen, 2025; K & K, 2026). One recent systematic review of machine learning for recidivism prediction found that the performance of ML models for this task was average 0.81 for accuracy and 0.74 for area under the curve on replicable datasets but noted that both the transparent algorithm or XAI approach and high-quality input data were necessary conditions for routine professional application [25]. Explainability-constrained sentencing models have been found to possess predictive validity not inferior to that of their opaque counterparts (0.71 versus 0.70-0.72 AUC on the ProPublica COMPAS dataset, $n = 6,172$) and meet due process standards [23].

However, no existing approach encompasses all of the following aspects which, combined, are critical for ID assessment: (i) multimodality of information due to the manifestation of risk through clinical, behavioural, cognitive and narrative information; (ii) time dynamics due to variability of risk based on changes in therapy progress and environment; and (iii) explainability as an inherent requirement due to both potential harm to a vulnerable group in the event of misclassification, and due to decision accountability of practitioners. This work aims to fill this gap through the development of a specialized multimodal risk prediction framework for ID based on XAI.

The contributions of this paper are as follows:

- Architecture of a multi-modal ID-specific risk and rehabilitation prediction system, which includes modality-specific encoders, adaptive fusion layer, and self-attention trajectory learning module.
- Three-level explainability approach (additive attribution, Shapley-based attribution, and counterfactual explanations) that allows for interpretability of the predictions for clinicians, forensic experts, and judges.
- Methodological approach to addressing class imbalance, fairness, and out-of-distribution issues based on previous evidence on class imbalance and algorithmic bias in clinical research.
- Evidence synthesis of the existing evidence on predictors of criminal risk and rehabilitation, with mapping to the evaluation framework of the proposed approach.

The rest of the paper is organized as follows. In Section 2, related works from four domains will be discussed: XAI for offending prediction, offending and rehabilitation of individuals with ID, multimodal data fusion, and explainability techniques. An explicit research gap will be stated at the end of this section. The proposed framework will be presented in Section 3. Section 4 outlines the methodology, while Section 5 provides the results of evidence synthesis along with tables, figures, and pipeline diagram.

Related Work

Explainable AI in Offending and Recidivism Prediction

Machine learning has become an established technology for predicting recidivism, and numerous assessments of the relative performance of ML versus traditional actuarial approaches have been conducted. According to a systematic review screening 79 papers and keeping 12 replicable studies, machine learning has achieved consistently good results, with accuracy averaging at 0.81 and average AUC of 0.74, and concluded that transparent models or the use of XAI technologies is one of the enabling factors for the professional adoption [25]. In a comparison of supervised learning algorithms, it has also been discovered that ensemble methods perform better on average and achieve higher predictive validity compared to older ones, where LogitBoost, random forests, and MultiBoosting are some of the most effective algorithms, but their advantage remains moderate, and no particular approach prevails in all cases [5].

XAI has become the main principle for the more modern studies. A behaviour-aware XAI system for criminal offender profiling was developed using behavioral features extracted from previous criminal offense reports and using ensemble-based ML models, and subsequently applying SHAP and LIME to generate both localized explanations for the predictions and global perceptions of behavioral importance.

In an empirical analysis on an existing data set of a correctional facility in an Eastern European nation, XGBoost was shown to have the best predictive accuracy (AUC = 0.83), where the factors influencing the re-offending behavior through SHAP analysis were determined to be age at release, offence committed before, level of education, and employment status [3]. The research on explainable and fair recidivism prediction revealed that some explainable ML models are capable of predicting recidivism as good as the black-box models and better than the COMPAS and Arnold PSA, and also, the models need to be created separately for each distinct geographical location and continuously updated [27].

Additionally, there is emerging research on the temporal and structural extension of this topic. In a research based on 20,457 records of in-prison offenders stored in a data management system of a prison in Southwestern China, the models were created for 1-, 2-, and 5-year windows of recidivism prediction and SHAP was used to discover feature contributions, where young offenders charged with larceny, with previous conviction, lesser fines, and less family ties were determined to have higher chance of being recidivist [18]. With regard to the governance level, the framework of AI-supported sentencing incorporated explainability as a basic constraint by means of a three-level explanation approach consisting of inherent interpretability through generalized additive models, exact local feature attribution obtained via additive model decomposition, and counterfactual reasoning involving the smallest possible input perturbations for changing risk classes [23]. This research tradition thus confirms not only the possibility but also the normative necessity of explainable offending predictions, although the former has been established almost exclusively with general offender populations.

Offending Behaviour and Rehabilitation in Individuals with Intellectual Disabilities

Epidemiology of ID among the forensic system is rather impressive and yet methodologically controversial. Prevalence of ID varies with differences in diagnosis and criminal justice process and varies internationally in terms of methodology between 2% and 40% [13]. Prevalence of ID among the prison population is estimated at 1% to 2% and below 5% in all cases. People with ID within the forensic population are usually younger than the rest of the population and their most frequent types of crimes are either violent and sexual or arson offenses [19]. Prevalence rates among forensic psychiatric units differ greatly and in a certain study it was 13.1% against 9.2% in a prior study in Ontario forensic psychiatric unit and general population prevalence of approximately 1% [14]. It should be noted that according to a Dutch maximum-security study ID prevalence reached 40%. Based on community-interface estimates, it is believed that approximately one in ten individuals with ID will interact with the police or judicial system in terms of committing criminal offenses, while males with ID are three times and females are four times as likely to have been convicted in the past as compared to those individuals without an intellectual disability, with the difference being even greater for violent criminal offenses – four times higher among males and 25 times among females [8].

There seems to be limited literature concerning risk factors for individuals with ID. A study conducted on 145 offenders with ID who were discharged from four independent sector medium secure units in the UK between 1990 and 2001 showed that criminal history variables were significantly associated with general recidivism, and that the Offender Group Re-conviction Scale – a criminogenic tool used to assess general offenders – showed great predictive validity among offenders with ID [7]. It is a two-sided sword as it means that criminogenic tools are generalizable to other populations, but it also implies lack of ID-specific tools.

In terms of rehabilitation, research has found that adaptive interventions can be effective. A twenty-year study of a community forensic intellectual disability service found that while the recidivism rates are poor, harm reduction results have shown that assessment and treatment for offenders with ID is highly effective from the perspective of public safety [16]. Comparison between mainstream versus specialized cognitive behavioral therapy programs found a "responsivity threshold," where mainstream services provide more inclusive society but use non-adaptive cognitive protocols that exceed the processing capacities of offenders with ID, thus increasing attrition and preventing skills generalization; specialized programs, especially adapted dialectical behavior therapy and specialist therapeutic communities, displayed higher engagement and a significant reduction in outward aggression, thus the authors suggested a hybrid approach that provides adaptations, like simplified visual prompts and changed pacing [4]. Long-term treatment outcome data provide evidence of both the potential and limitations of intervention; a follow-up study of men with ID who were involved in the sex offender treatment groups found that in the literature generally, 13.4% of men re-offended within four to five years, and a review of treatment services among 62 sex offenders with intellectual disabilities revealed that 4% re-offended within the first year and 21% after four years [9].

Multifactorial assessment has been proved effective for enhancing rehabilitation plans. According to a mixed methods study on the implementation of biopsychosocial multifactorial assessment model in three secure forensic units (N = 120 offenders with mild and moderate ID; mean age: 29.4 ± 7.8 years; 82% males), 71% had a concurrent psychiatric condition, 57% experienced substantial adverse childhood experiences, and 49% used drugs; after assessment using the biopsychosocial model, there was a 35% increase in individual treatment plans, 28% increase in risk management strategies, and a 22% decrease in incidents of misbehaviour ($p < 0.05$); multiple regression analysis has revealed that experience of trauma ($\beta = 0.41$) and substance misuse ($\beta = 0.36$) are associated with institutional aggression [6]. Specific fourth-generation risk prediction models based on the concepts of dynamic risk management and rehabilitation have been developed for the problem of sexual offending in individuals with ID by working with a proxy group (N = 25,000) extracted from the NIJ Recidivism Forecasting Challenge dataset, where the baseline three-year re-offending rates for sexual offenses ranged from 4% to 8% among high-risk offenders and high program attendance (>10 sessions) was associated with a 15-20% lower likelihood of re-offense (OR = 0.82) [21].

Multimodal Artificial Intelligence and Data Fusion

The rationale for multimodal modelling lies in the clinical observation that clinicians do not reason based on a single data source. As noted in a systematic review and implementation guide about combining medical imaging and electronic health records, state-of-the-art radiology models focusing on pixel-value information fail to incorporate the clinical context needed for proper interpretation of results, and thus fusion approaches were recommended for deep learning models incorporating clinical context along with pixel information [11]. In a scoping review of 34 studies, it was established that fusion of imaging data with EHR is gaining traction, doubling from 2020 to 2021; that early fusion was the most popular approach in majority of applications (22 out of 34 studies); and that fusion-based multimodality models performed better than traditional single-modality models for the same tasks, disease diagnosis and prediction being the most popular outcomes [17]. A more extensive literature review about machine learning and deep learning-based multimodal architectures concluded that multimodal approaches generally outperform unimodal approaches regarding diagnostic accuracy, robustness, and generalisation, and identified different types of fusion approaches and architectures including hybrid architectures, transformers for vision-language models, and EHR-based approaches [12]. Another survey on multimodal machine learning in healthcare presented a survey of 17 multimodal clinical datasets and over 50 papers – mainly concerning image-tabular integration – and found that although multimodal approaches can help to increase prediction accuracy, the usefulness of a particular approach depends on the data and task at hand [15]. With regard to the ID

risk-assessment problem, however, the literature provides a set of tools, but no applications. The problem of ID involves the combination of cognitive tests, behavioral sequences, and criminological history.

Explainability Methods

SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), the two most extensively employed model-agnostic methods, fill complementary niches. SHAP is based on game theory, and it delivers theoretically well-supported importance values of the features that show their impact on each particular prediction and can thus be summed up to deliver global importance values for all predictions; LIME, on the contrary, employs simple interpretable models to explain individual predictions, doing so locally around a particular observation [22]. Such a complementary scheme of global insights from SHAP and local decision from LIME explains why they are combined in clinical studies so often [22].

Nevertheless, there is also some criticism in the literature on such an approach. There was even a special perspective article devoted to the application of SHAP and LIME in biomedical studies, which claimed that both the methods were quite sensitive to the choice of ML model and feature collinearity [22]. This is directly applicable in ID risk assessment, where clinical markers such as adaptive function and intelligence quotient, as well as psychiatric co-morbidity and medications, tend to be naturally associated and, therefore, may lead to incorrect attributions of significance.

Ethics, Fairness, and Algorithmic Bias

Risk assessment for criminal justice represents the classic case study of algorithmic discrimination. The controversy over COMPAS has shown that even investigative journalism can unveil systematic differences – there was found to be the possibility of racial discrimination during the use of the algorithm for over 18,000 criminal defendants in Broward County, Florida [26]. The mechanism through which the algorithm may cause harm goes beyond the issue of static bias and can create a vicious cycle: when one of the assessment criteria is based on the criminal record of the defendant's peers, the defendant is seen as more likely to commit crimes due to his associations, thus being sentenced to jail, which increases the amount of criminality within a particular community only based on correlation [26]. This situation is particularly dangerous for those with ID, as their social network can be limited not by choice, but by provision of services and institutional care.

Research Gap

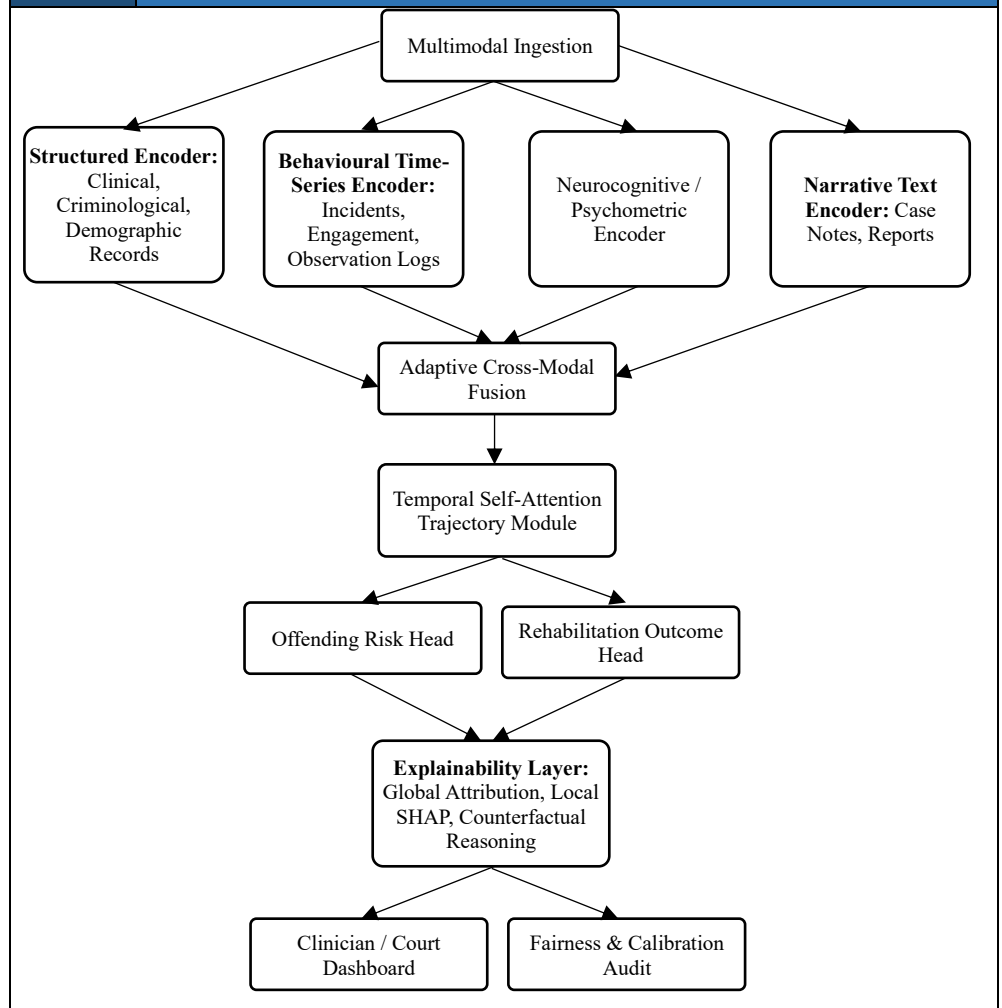
Based on the strands listed above, there emerges an agreement regarding three unfulfilled needs. First, the body of research related to offending predictions is very advanced technologically yet has been mostly conducted on general or non-ID sample populations even though there is evidence showing that ID has different dynamics (Bhattacharya & Dora, 2025; Jones, 2007). Secondly, the multimodal fusion literature shows that fusion increases the accuracy and robustness of models but has not been used in the prediction of risks and rehabilitation of ID offenders [15] Mohsen et al., 2022). Thirdly, explainability has been identified as necessary for professional adoption [25], although it is often considered a post hoc need. There are issues with the sensitivity of explainable methods to models and collinearity [22] and the risks posed by fairness in high-stakes justice scenarios [26].

Proposed Framework

Overall Architecture

The architecture consists of five layers: (i) multimodal ingestion, (ii) modality-specific encoding, (iii) adaptive fusion, (iv) temporal trajectory modeling, and (v) dual-headed prediction block connected with the explainability layer in figure. 1. The risk prediction and rehabilitation predictions are made in a joint manner, utilizing the already known correlation between dynamic risk management and rehabilitation [21].

Figure 1 Pipeline of Proposed Multimodal Risk Assessment Framework Utilizing Xai Technology



Modalities and Feature Domains

Table 1	Modality-feature mapping in the suggested framework		
Modality	Data type	Representative features	Clinical rationale
Structured records	Tabular	Age at Index Offence, Prior Convictions, Index Offence Type, Comorbid Psychiatric Diagnosis, Substance Misuse History, Trauma Exposure, Level Of Supervision	Static and Historical Criminogenic and Clinical Variables [6][7]
Behavioural time-series	Sequential	Programme Attendance, Incident Frequency, Mood/Stress Ratings, Social Interaction Indices, Environmental Stressor Events	Dynamic Risk Factors That Fluctuate and Mediate Escalation [20]
Neurocognitive / psychometric	Tabular	Cognitive Functioning, Adaptive Behaviour Scores, Impulse Control Measures, Psychological Test Results	Neurodevelopment-Specific Responsivity and Capacity [2]
Narrative text	Unstructured	Case Notes, Behavioural Reports, Risk Formulation Narratives, Treatment Summaries	Contextual, Idiographic Detail not Captured Numerically

Four modalities are fed into this framework, selected in a way that corresponds to the biopsychosocial factors identified as significant among forensic identification population [6], as well as to the dynamic risk factors that are insufficiently covered by traditional measures [20]. Table 1 provides this correspondence for all modalities.

Adaptive Fusion

The choice of fusion strategy is important for architecture design. The multimodal literature shows that early fusion has been the most widely used approach in applications and that fusion models perform better than unimodal models [17], whereas other reviews state that the best choice depends on both task and data characteristics [15]. For this reason, the

proposed framework utilizes adaptive attention-gated late-intermediate fusion, whereby encoders process modality data into embeddings that get fused via attention weights, thus giving more weight to useful information for each particular subject instead of always using concatenation. Such a design pattern resembles attention-gated hybrid fusion architectures that have shown themselves to be successful when dealing with heterogeneous and variable modalities [12]. This is especially relevant in our case since the availability of data itself can convey useful information – a behavioral modality poorly observed can mean disengagement.

Temporal Trajectory Modelling

The risk in this population is non-stationary, and thus the output of the fusion layer is passed to a trajectory module using temporal self-attention. Transformers with self-attention have been proven capable of capturing long-distance dependencies in time-series clinical data without recursion and hence without problems associated with vanishing gradients, as well as being able to perform parallel computations [24]. Sparse and irregular observations can be handled by observation triplet representation and value embedding in the trajectory module; this is essential due to the irregular nature of forensic behavioural data [24]. This module generates both the current state risk prediction and a prediction of the rehabilitation trajectory, allowing the clinician to see where the patient currently is and where he/she will go; a feature in tandem with the fourth-generation logic of dynamic-risk and rehabilitation orientation.

Explainability Layer

Explainability operates at three levels according to a multi-level architecture approach [23] but suited to multimodal outputs:

1. Aggregate attribution. Global attribution using SHAP explains which attributes or modalities influence the prediction of the cohort, and helps with audit and review at the service level [22].
2. Local attribution. Individual attributions using SHAP and LIME explain individualized predictions, where SHAP provides additive contributions theoretically supported by the approach, while LIME uses a local fit of a simpler model [22].
3. Counterfactual generation. Changes in the input needed to change the risk classification are calculated and interpreted as interventions for the individual (e.g., how much programme attendance should be increased to shift the person from one classification into another) [23].

Since the reliability of explanations depends on model architecture and feature collinearity [22], the approach combines each explanation with a collinearity and stability audit. This prevents the system from presenting spurious certainty in the correlated clinical domains characteristic of ID assessment (e.g., cognition and adaptive functioning).

Methodology

Cohort and Data Sources

The research framework is intended for use in a retrospective, multi-site forensic ID cohort developed in forensic units and community forensic services with future validation serving as the second phase. Evidence supporting this cohort includes the findings from mixed-method forensic investigations of similar magnitude, such as a three-unit investigation involving $n = 120$ offenders who have mild to moderate ID (mean age 29.4 ± 7.8 years; 82% males) [6], and the registry-level studies like the dataset of 20,457 prisoners used for time window recidivism modelling [18]. Criteria for inclusion include an ID diagnosis, an index offense, or any other forensic contact, and at least one follow-up assessment, whereas criteria for exclusion include incomplete risk documentation and absence of any behavioral observation.

Preprocessing and Feature Engineering

Preprocessing is done modality by modality because each modality needs separate harmonization. The structured and psychometric data are standardized along with filling in for missing values; the behavioural time series is resampled into the same cadence and represented as observation triplets to retain irregularity [24]; the narrative text is anonymized and represented using a domain-adapted language encoder. Temporal alignment is

performed across modalities, where the common index date and future assessments are the reference dates for all the features (this is necessary to prevent information about the future from leaking into any of the features) [27].

Handling Class Imbalance

Reoffending and rehabilitation failure instances occur much less frequently compared to non-instances, and skewed datasets have a tendency to bias classifiers against the minority class and misclassify it, which in this case will be the class of interest [28]. The method used here utilizes a two-tier approach, namely, (i) data-level resampling (random oversampling of the minority class and/or hybrid oversampling) and (ii) cost-sensitive learning, i.e., adjusting the cost function for the class weights to inverse class frequencies. The first two methods have been empirically validated on imbalanced clinical datasets: while random oversampling and undersampling are popular practices, the influence on discrimination varies [28]; cost-sensitive learning is able to compensate for imbalances without losing majority-class information. The model choice is based on precision–recall and calibration statistics rather than accuracy, because accuracy is a misleading measure for the base rates of reoffending.

Models and Training

To give both a clear point of comparison and a high-throughput comparator, the benchmarking framework compares the multimodal model to:

- Logistic regression — a clear interpretive linear baseline for effect-size estimation and clinical communications;
- Ensemble learners (random forests, gradient boosting, and XGBoost) — reliable performers in recidivism predictions [5]; Travaini et al., 2022);
- Generalized additive models — intrinsically interpretable architectures for accurate local attribution [23];
- Temporal transformers — the principal architecture for temporal fusion [24].

The models are trained using stratified and temporally ordered split, which involves training on early samples and testing on later samples because of the conclusion that different regions require different models and they must evolve over time [27]. The hyperparameters are fine-tuned using a validation partition.

Evaluation Metrics

The evaluation follows a multi-criterion approach in line with the existing literature on recidivism prediction [25]: area under the receiver operating characteristic curve and area under the precision-recall curve for discrimination; sensitivity, specificity, and positive predictive value for decision thresholds; Brier score and calibration plots for probabilistic accuracy; and subgroup criteria (by gender, level of ID, and index offence category) for fairness. The framework provides decision thresholds explicitly, since risk instruments are actually employed as probabilities, not binary classifiers [23].

Explainability Evaluation

The quality of explanations is measured using three dimensions, which include fidelity (alignment between the significance attribution and behavior of the model during perturbation), stability (variability in explanations when resampling and under feature correlation [22], and human value (clinicians' structured assessments on whether the explanations help make the decisions or just follow the decision process). For counterfactual explanations, actionability, which measures the feasibility of suggested modifications, is used.

Ethical Governance and Fairness Auditing

In consideration of the susceptibility of the people involved, along with the known issues associated with algorithmic bias in the justice system [26], these governance criteria have been defined: compulsory testing for unfairness with specified disparity limits; the banning of use of features that would lead to self-reinforcing loops based on correlated information such as criminal history of associates [26]; human-in-the-loop decision-making, where the model only aids but does not control the decisions; and procedural-justice level transparency [23].

Results and Evidence Synthesis

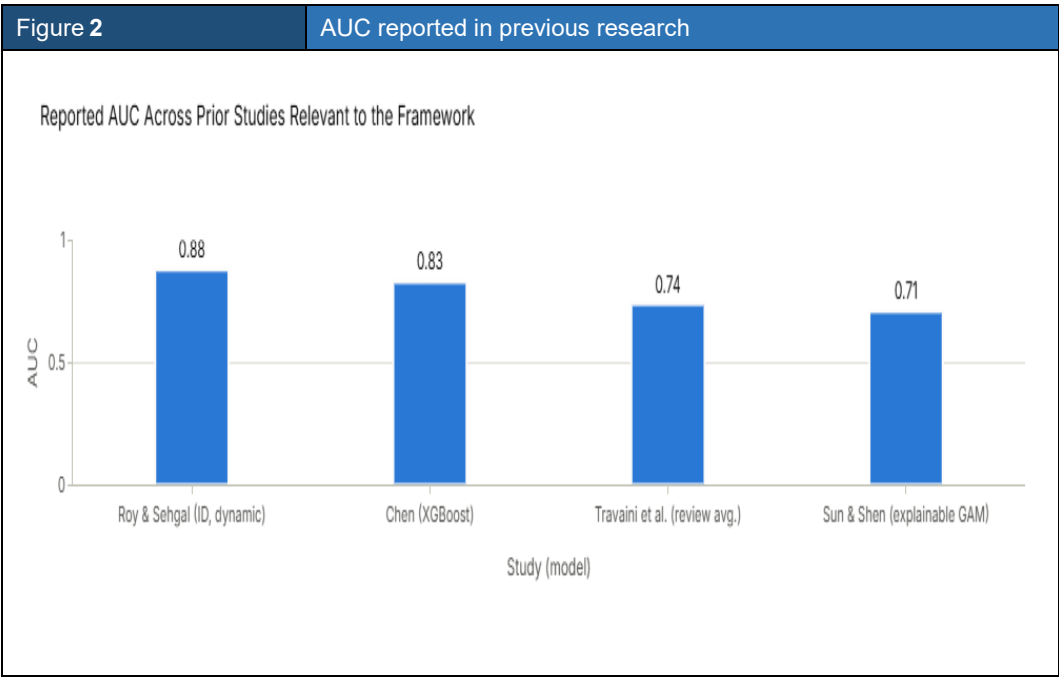
Since the proposed framework was not implemented in the context of a novel sample, two types of result are given in this section. First, it summarizes the empirical results as described in previous work so as to determine the feasible range of performance and the standard for evaluating the proposed framework. Second, the evaluation strategy for the proposed framework is detailed along with its potential strengths. No numbers that have been generated synthetically or through simulation are given in the context of the results of the proposed framework; all numeric data below have come directly from the source material.

Predictive Performance Across Prior Studies

Table 2 provides a summary of the performance measures for selected models in each of the offending prediction and ID-risk literature. Noteworthy is the variability, with models trained on general offender groups being grouped into the range of AUC 0.70 to 0.83, whereas the single ID-focused dynamic model included exhibits a considerably larger AUC.

Table 2	Predictive Performance of Models Associated with the Framework		
Study (Population)	Method	Key Discrimination Metric	Other Reported Metrics
Roy & Sehgal (Offenders with ID, n = 300) [20]	Logistic Regression, Decision Trees, Random Forests (Dynamic Model)	AUC = 0.88 (Dynamic) vs 0.75 (Static)	Sensitivity 85%; Specificity 78%; Accuracy 72%→82% (~30% Improvement)
Chen (Correctional Sample, Eastern Europe) [3]	Logistic Regression, Random Forest, XGBoost	AUC = 0.83	SHAP-Identified Predictors: Age at Release, Prior Offence Type, Education, Employment
Travaini et al. (Systematic review, 12 Studies) [25]	Multiple ML Techniques	Average AUC = 0.74	Average Accuracy = 0.81
Sun & Shen (COMPAS, n = 6,172) [23]	Explainability-Constrained GAM (GA2M)	AUC = 0.71 vs 0.70–0.72 (black-box)	Comparable Validity with Constitutional Due Process Compliance

The relative picture is visualised in figure 2.



There are three points to be made in connection with this data. The first one is that the explainability-constrained models demonstrate a level of discrimination statistically comparable to that of the opaque models [23], which immediately refutes the assumption that transparent models necessarily have to give up on accuracy. The second point is that the maximum AUC value within this set of models has been demonstrated by an ID-specific dynamic model [20], and thus the framework is right to prioritize dynamic and contextual

characteristics. Finally, in terms of the review, AUC = 0.74 indicates the uncertainty of the problem.

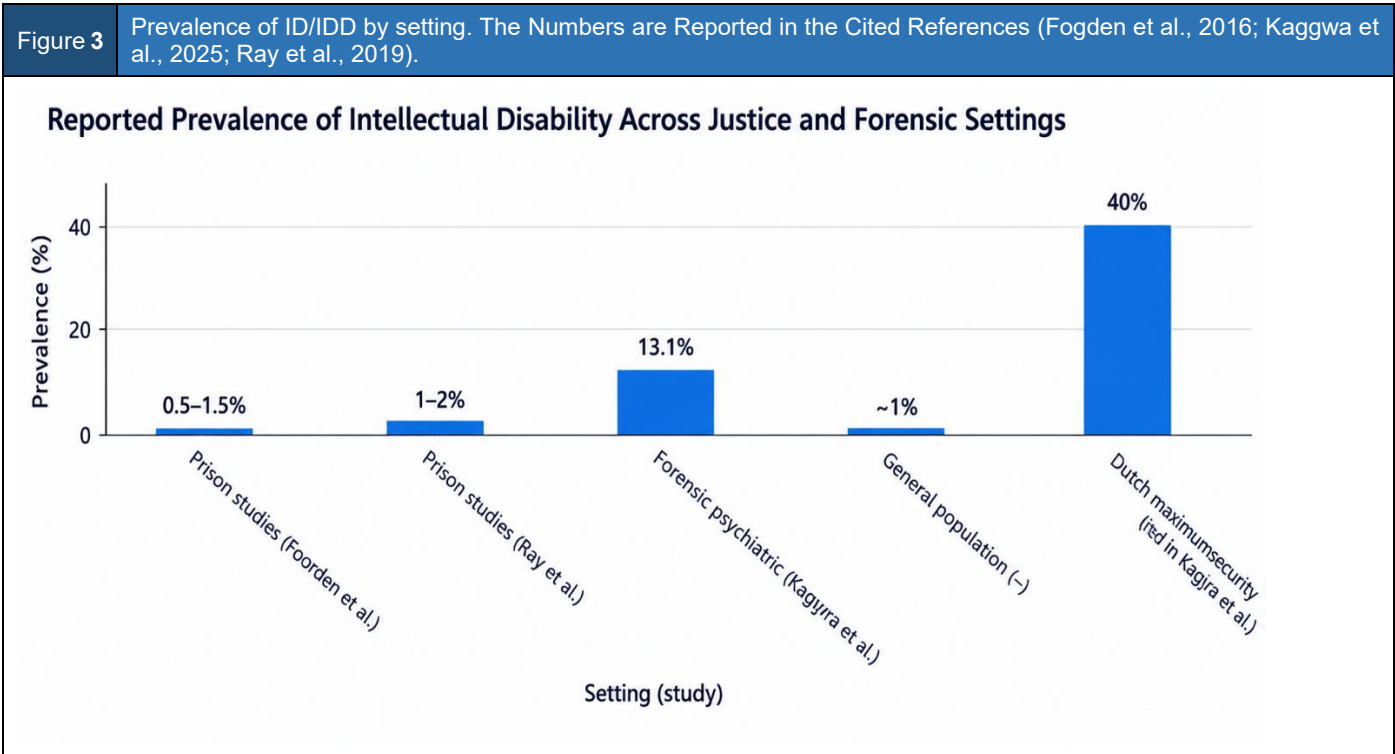
Predictor Importance and Effect Sizes

Table 3 shows the predictors that have been noted and their effect sizes where known. The parallel between the criminogenic factors (such as previous convictions and age) and the clinical/dynamic factors (like trauma, substance abuse, and program participation) provides the evidence for multimodal integration.

Table 3		Identified Factors Influencing Offending and Rehabilitation-Related Outcomes Among ID Offenders		
Predictor	Direction	Reported Effect		Source
Trauma exposure	↑ Institutional Aggression Risk	$\beta = 0.41$ (Multivariate Regression, $p < 0.05$)		[6]
Substance misuse	↑ Institutional Aggression Risk	$\beta = 0.36$		[6]
Programme attendance (>10 sessions)	↓ Recidivism Odds	OR = 0.82 (15–20% Lower Odds)		[21]
Prior convictions, age at release	Dynamic Moderators Of Outcome Variance	Identified as Stable Baseline Factors		[21]
Age at Release, Prior Offence Type, Education, Employment	Key Predictors Of Re-Offending	SHAP-Identified		[3]
Criminal History Variables	↑ General Recidivism	Significant at 2-Year Follow-Up		[7]

Prevalence and Rehabilitation Context

The need for a framework that is specific to ID is underscored by the statistics shown in figure 3.



Evidentiary basis for rehabilitation outcomes reinforces the two-tiered nature of the proposed model. For instance, in one community sample of 103 men with ID who were under treatment, 10.7% reoffended sexually after a median follow-up of 5.8 years, and majority of such offenses were non-contact and involved victims familiar to the perpetrator [1]. In a meta-analysis of group-based CBT program for men with ID who were offenders in terms of sexual behavior, significant decreases in cognitive distortions, increases in sexual knowledge and victim empathy were found along with an incidence rate of 11.5% of sexually abusive behaviors reported in the follow-up period compared to 10% among treatment completers without ID and 13.7%-14.1% among untreated comparators in the cited studies [10]. Structured biopsychosocial assessment led to 35% increases in augmented care plans and 22% decreases in reported conduct episodes (Dwivedy & Yadav, 2025).

Framework Evaluation Design and Expected Capabilities

Evaluation Plan in table 4 is specified by concrete goals, rather than results. It is important to note that no performance statistics are mentioned for the suggested system, and the plan is such that any future research may prove each particular row true or false.

Table 4	Evaluation plan and target capabilities of the framework	
Dimension	Evaluation approach	Benchmark context from prior work
Discrimination (Offending)	Temporal Hold-Out AUC and AUPRC	Prior AUCs 0.71–0.88 [20][23]
Discrimination (Rehabilitation)	Regression/Classification Metrics on Outcome Vector	Outcome rates 10.7%–11.5% re-offending [1][10]
Calibration	Brier Score, Calibration Curves	Probability-score usage expected [23]
Fairness	Subgroup AUC/Sensitivity and Disparity Thresholds	Bias hazards documented [26]
Explainability fidelity	Perturbation Agreement, Explanation Stability	SHAP/LIME sensitivity to model and collinearity [22]
Clinical utility	Structured Clinician Ratings, Counterfactual Actionability	Human-facing explanation requirement
Robustness	Location-Wise and Temporal Retraining	Location-specific training advised [27]

Discussion

The key point being made by this paper is that explainability, multimodality, and fairness are not something that should be added to an ID risk assessment in sequence, but rather are equally important constraints for the design because of the population and decision context. Evidence synthesis provides an argument from the empirical perspective but also demonstrates the tensions in implementation.

Explainability does not entail a loss of accuracy. The usual argument against transparency of models is that interpretability comes at the cost of performance. However, an explainability-constrained framework based on COMPAS managed to achieve AUC 0.71 compared to 0.70–0.72 for black boxes while meeting the requirements of due process [23]. It has also been shown that interpretable models were able to predict recidivism just as accurately as black box models, and even outperformed COMPAS or the Arnold PSA [27]. This is especially important in the case of ID since if explanation entails no loss of predictive power, there is absolutely no justification for using a black box model over the vulnerable population.

Multimodality is empirically supported, not just an architectural choice. Predictors found in the different studies include: Variables related to criminal history predicted recidivism at two-year follow-up in individuals with intellectual disabilities (Fitzgerald et al., 2011); trauma exposure ($\beta = 0.41$) and substance abuse ($\beta = 0.36$) predicted institutional aggression (Dwivedy & Yadav, 2025); and programme participation was negatively related to recidivism (OR = 0.82) (Sahu & Tanwani, 2025). None of the modalities encapsulates these cues. Meta-analyses uniformly confirm that fusion models are superior to unimodal models and enhance robustness and generalisation (Jandoubi & Akhloufi, 2025; Mohsen et al., 2022), which is the technical equivalent of this complementary relationship. Adaptive fusion in the framework is therefore justified not by fad architecture but by the nature of the risk construct itself.

The use of dynamic modelling accounts for an instrument malfunction documented in research. Traditional instruments are inefficient in response to dynamic risk factors because of changing situations of ID offenders, which have an impact on their behavior [20]. It is evident from the fact that ID-specific dynamic model performed significantly better than the static one (AUC 0.88 vs 0.75; sensitivity 85%; specificity 78%) [6], proving the benefit of time-related and dynamic characteristics which drive the design of time-series encoder and temporal attention modules in the proposed framework [24]. Fourth generation rehabilitation-oriented frameworks follow the same approach measuring protective factors and responsivity instead of applying high/low cut-offs (Sahu & Tanwani, 2025).

Rehabilitation being mutable and historical risk factors not, a predictor framework that is able to predict both is more potent clinically than one that only predicts risk (Sahu & Tanwani, 2025).

Fairness is not an issue that should be brushed aside lightly. The known issues of COMPAS, including detection of potential racial biases among more than 18,000 defendants [26], and

the self-reinforcing vicious cycle created by including associate criminal history in a risk prediction model [26], make clear how an otherwise accurate system could be inherently structurally biased. With people with ID already having an increased incidence of exposure to the justice system – three times as many males and four times as many females having a previous conviction, and significantly increased odds for violent crimes [8] – fairness auditing and the exclusion of feature selection driven by correlation become an ethical requirement.

A rigorous methodology is an absolute necessity. The predictiveness of recidivism prediction is still suboptimal, the average AUC being 0.74 among studies that can be replicated [25], with models requiring location-specific training and periodic updating [27]. Re-offending is a rare phenomenon that skews algorithms in favor of the majority group [28]. Thus, validation of the framework in terms of precision-recall and calibration becomes necessary rather than accuracy. Lastly, as SHAP and LIME techniques are sensitive to choice of model and collinear features [22], explanation stability auditing is critical when cognitive and adaptive assessment is necessarily correlated.

In combination, all these arguments lead to the following conclusion: while a multimodal and temporally conscious framework is possible and reasonable for ID offenders, its real value would not lie in discrimination but in calibration and fairness, as well as in how informative its explanations are for making humane decisions.

Limitations

There are several limitations to this study. Firstly, the proposed framework has not been empirically tested using a primary dataset, hence the performance values stated herein are from earlier research and cannot be assumed as output from the proposed model. Secondly, the prevalence and outcome values in the literature of offenders identification range broadly between 2% and 40%, based on the methodology used [13]; moreover, the estimated prevalence ranges from prison estimates of 1% to 2% and forensic psychiatric values of 13.1% (Kaggwa et al., 2025; Ray et al., 2019), thus composition of the cohort will highly affect the evaluation process. Thirdly, the risk factors associated with identification of offenders have largely been identified based on relatively small sample sizes and retrospective studies and the field has been plagued by inconsistencies in definitions and methodologies [13]. Fourthly, the explanation methods themselves have reliability constraints due to dependence on model architecture and features collinearity [22]. Fifthly, fairness assurances in this case are merely procedural, and the disparities among subgroups will remain despite audits due to lack of data.

Conclusion

In this paper, the development of a new explainable artificial intelligence-based framework was proposed, which can help with the prediction of offending behavior and rehabilitation outcomes in individuals with intellectual disabilities. It includes the integration of structured data from clinical and criminological records, time series behavioral information, neuropsychometric assessments, and the textual narrative of cases with modality-specific encoders and cross-modal fusion, which takes into account the longitudinal nature of the data using a temporal self-attention module. In the proposed framework, explainability is incorporated on multiple levels: global attribution, local explanation per specific case and counterfactual reasoning, and it is combined with stability auditing, which will help to avoid spurious certainty related to correlated clinical modalities. Issues of fairness, dealing with class imbalance and temporal evaluation were considered as essential aspects of methodology and not as additional components. The integration of available evidence suggests that explainable models could perform similarly to their opaque counterparts in terms of discrimination, that dynamic and multi-faceted features have a lot of informative content in ID offender population and that rehabilitation outcomes are measurable and modifiable. Further research should be done on prospective validation at multiple sites, replication across jurisdictions, and clinician assessment of explanation utility.

References

1. Barros, Sofia, Claudia Oliveira, Eduardo Araujo, Diana Moreira, Fernando Almeida, and Anita Santos. "Community intervention programs for sex offenders: A systematic review." *Frontiers in psychiatry* 13 (2022): 949899. <https://doi.org/10.3389/fpsyt.2022.949899>

2. Bhattacharya, Akash, and Aida Ventkat Rao Dora. "PRISMA ScR Guided Scoping Review of Forensic Models of Care in Intellectual Disability." *Journal of Intellectual Disabilities and Offending Behaviour* 16, no. 3 (2025): 97-104. <https://doi.org/10.47059/jidob/v16/i3/11>
3. Chen, Tianxin. "Predicting Recidivism Risk Using Explainable Machine Learning Models: A Case Study from Eastern European Correctional Data." *Journal of Criminal Investigation and Criminology* 76, no. 2 (2025). <https://doi.org/10.52152/nh8nfx77>
4. Dhokia, Gautam, Annapurna Pattnaik, Teena Kiran Nichant, Uma Bhardwaj, Kasturi Pohini, and Aashim Dhawan. "Comparative Evaluation of Mainstream Versus Specialist CBT Programmes for Offenders with Intellectual Disabilities." *Journal of Intellectual Disabilities and Offending Behaviour* 16, no. 4 (2025): 1-7. <https://doi.org/10.47059/jidob/v16/i4/1>
5. Duwe, Grant, and KiDeuk Kim. "Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism." *Criminal Justice Policy Review* 28, no. 6 (2017): 570-600. <https://doi.org/10.1177/0887403415604899>
6. Dwivedy, Dilip, and Suresh Chandra Yadav. "Biopsychosocial Assessment Approach to Offenders with Intellectual Disabilities in Forensic Services." *Journal of Intellectual Disabilities and Offending Behaviour* 16, no. 2 (2025): 10-20. <https://doi.org/10.47059/jidob/v16/i2/2>
7. Fitzgerald, Suzanne, Nicola S. Gray, John Taylor, and Robert J. Snowden. "Risk factors for recidivism in offenders with intellectual disabilities." *Psychology, Crime & Law* 17, no. 1 (2011): 43-58. <https://doi.org/10.1080/10683160903392293>
8. Fogden, Billy C., Stuart DM Thomas, Michael Daffern, and James RP Ogloff. "Crime and victimisation in people with intellectual disability: a case linkage study." *BMC psychiatry* 16, no. 1 (2016): 170. <https://doi.org/10.1186/s12888-016-0869-7>
9. Heaton, K.M. and Murphy, G.H., 2013. Men with intellectual disabilities who have attended sex offender treatment groups: A follow-up. *Journal of applied research in intellectual disabilities*, 26(5), pp.489-500. <https://doi.org/10.1111/jar.12038>
10. Heppell, Stacey, Christopher Jones, and John Rose. "The effectiveness of cognitive-behavioural therapy group-based interventions for men with intellectual disabilities and sexual offending histories: a meta-analysis." *International journal of developmental disabilities* 68, no. 4 (2022): 416-429. <https://doi.org/10.1080/20473869.2020.1845933>
11. Huang, Shih-Cheng, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines." *NPJ digital medicine* 3, no. 1 (2020): 136. <https://doi.org/10.1038/s41746-020-00341-z>
12. Jandoubi, Bassem, and Moulay A. Akhloufi. "Multimodal artificial intelligence in medical diagnostics." *Information* 16, no. 7 (2025): 591. <https://doi.org/10.3390/info16070591>
13. Jones, Jessica. "Persons with intellectual disabilities in the criminal justice system: Review of issues." *International journal of offender therapy and comparative criminology* 51, no. 6 (2007): 723-733. <https://doi.org/10.1177/0306624x07299343>
14. Kaggwa, Mark Mohan, Joan Abaaty, Arianna Davids, Luke Brenton, Madeline Komar, John MW Bradford, Gary Andrew Chaimowitz, and Andrew Toyin Olagunju. "Self-harming behaviors among forensic psychiatric patients living with intellectual disability." *BMC psychiatry* 25, no. 1 (2025): 976. <https://doi.org/10.1186/s12888-025-07431-w>
15. Krones, Felix, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. "Review of multimodal machine learning approaches in healthcare." *Information Fusion* 114 (2025): 102690. <https://doi.org/10.1016/j.inffus.2024.102690>
16. Lindsay, William R., Lesley Steptoe, Lisa Wallace, Fabian Haut, and Eleanor Brewster. "An evaluation and 20-year follow-up of a community forensic intellectual disability service." *Criminal Behaviour and Mental Health* 23, no. 2 (2013): 138-149. <https://doi.org/10.1002/cbm.1859>
17. Mohsen, Farida, Hazrat Ali, Nady El Hajj, and Zubair Shah. "Artificial intelligence-based methods for fusion of electronic health records and imaging data." *Scientific Reports* 12, no. 1 (2022): 17981. <https://doi.org/10.1038/s41598-022-22514-4>
18. Mu, Di, Simai Zhang, Ting Zhu, Yong Zhou, and Wei Zhang. "Prediction of recidivism and detection of risk factors under different time windows using machine learning techniques." *Social*

19. Ray, Ipsita, Alexander IF Simpson, Roland M. Jones, Kristina Shatokhina, Anupam Thakur, and Benoit H. Mulsant. "Clinical, demographic, and criminal behavior characteristics of patients with intellectual disabilities in a Canadian forensic program." *Frontiers in Psychiatry* 10 (2019): 760. <https://doi.org/10.3389/fpsyt.2019.00760>
20. Roy, Jainish, and Rajesh Sehgal. "Dynamic Risk Assessment Model for Violence in Offenders with Intellectual Disability." *Journal of Intellectual Disabilities and Offending Behaviour* 16, no. 4 (2025): 44-54. <https://doi.org/10.47059/jidob/v16/i4/7>
21. Sahu, Ashish Kumar, and Shilpi Nishant Tanwani. "Fourth-Generation Risk Assessment Framework for Sexual Offending in People with Intellectual Disability." *Journal of Intellectual Disabilities and Offending Behaviour* 16, no. 3 (2025): 60-68. <https://doi.org/10.47059/jidob/v16/i3/7>
22. Salih, Ahmed M., Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz. "A perspective on explainable artificial intelligence methods: SHAP and LIME." *Advanced Intelligent Systems* 7, no. 1 (2025): 2400304. <https://doi.org/10.1002/aisy.202400304>
23. Sun, Jie, and Tao Shen. "AI-assisted sentencing modeling under explainability constraints: Framework design and judicial applicability analysis." *Information* 17, no. 3 (2026): 234. <https://doi.org/10.3390/info17030234>
24. Tipirneni, Sindhu, and Chandan K. Reddy. "Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, no. 6 (2022): 1-17. <https://doi.org/10.1145/3516367>
25. Travaini, Guido Vittorio, Federico Pacchioni, Silvia Bellumore, Marta Bosia, and Francesco De Micco. "Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction." *International journal of environmental research and public health* 19, no. 17 (2022): 10594. <https://doi.org/10.3390/ijerph191710594>
26. Tsamados, Andreas, Nikita Aggarwal, Josh Cows, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. "The ethics of algorithms: key problems and solutions." *AI & society* 37, no. 1 (2022): 215-230. <https://doi.org/10.1007/s00146-021-01154-8>
27. Wang, Caroline, Bin Han, Bhrij Patel, and Cynthia Rudin. "In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction." *Journal of Quantitative Criminology* 39, no. 2 (2023): 519-581. <https://doi.org/10.1007/s10940-022-09545-w>
28. Yang, Cynthia, Egill A. Fridgeirsson, Jan A. Kors, Jenna M. Reys, and Peter R. Rijnbeek. "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data." *Journal of Big Data* 11, no. 1 (2024): 7. <https://doi.org/10.1186/s40537-023-00857-7>