Applied Soft Computing Journal 176 (2025) 113160

Contents lists available at ScienceDirect

Applied Soft Computing

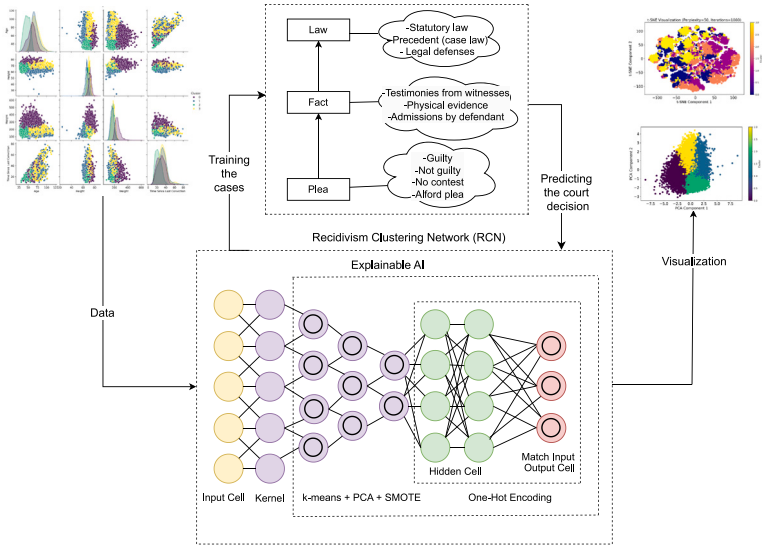
journal homepage: www.elsevier.com/locate/asoc

Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice

Muhammed Cavus [a,b,*], Muhammed Nurullah Benli [c,d], Usame Altuntas [e,f], Mahmut Sari [g], Huseyin Ayan [h,i], Yusuf Furkan Ugurluoglu [h,j]

[a] Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle upon Tyne, NE1 8SA, UK
[b] School of Engineering, Iskenderun Technical University, Iskenderun, 31200, Turkiye
[c] School of Law, University of Aberdeen, Aberdeen, AB24 3UB, UK
[d] School of Law, Ankara University, Ankara, 06590, Turkiye
[e] School of Law, Northumbria University, Newcastle Upon Tyne, NE2 1XA, UK
[f] Faculty of Law, Istanbul Medeniyet University, Istanbul, 34700, Turkiye
[g] Department of Construction, Kirsehir Ahi Evran University, Kirsehir, 40100, Turkiye
[h] School of Engineering, Newcastle University, Newcastle Upon Tyne, NE1 7RU, UK
[i] School of Engineering, Istanbul University-Cerrahpasa, Istanbul, 34320, Turkiye
[j] Department of Mechanical Engineering, Necmettin Erbakan University, Konya, 42090, Turkiye

GRAPHICAL ABSTRACT



ARTICLE INFO

Dataset link: https://github.com/cavusmuhammed68/Recidivism-Clustering-Network-RCN-

Keywords:
Deep learning

ABSTRACT

This paper presents the Recidivism Clustering Network (RCN), an effective approach for predicting repeat offenses using deep learning (DL), clustering, and explainable AI (XAI). The RCN improves offender profiling for more accurate and interpretable recidivism predictions, aligning with key legal principles like fair sentencing, transparency, and non-discrimination. The RCN employs machine learning (ML) models optimized with a

* Corresponding author at: Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle upon Tyne, NE1 8SA, UK.
E-mail address: muhammed.cavus@northumbria.ac.uk (M. Cavus).

https://doi.org/10.1016/j.asoc.2025.113160
Received 7 January 2025; Received in revised form 5 April 2025; Accepted 12 April 2025
Available online 29 April 2025

Recidivism prediction
Explainable AI
Criminal justice system

Keras tuner, using the Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance. With about 75% accuracy, the model shows strong recall, identifying 10,661 recidivists but producing 4,038 false positives—indicating a trade-off between sensitivity and specificity. Beyond predictions, RCN integrates clustering methods, including k-means, principal component analysis (PCA), and t-distributed Stochastic Neighbor Embedding (t-SNE), to identify hidden patterns within offender data. Visualizations reveal distinct clusters, linking characteristics, such as age, to recidivism behaviors. SHapley Additive exPlanations (SHAP) values enhance interpretability, showing that factors like time since the last conviction and age significantly impact predictions. The RCN approach offers substantial potential for criminal justice applications by combining predictive power with actionable insights, supporting a more ethical and accountable use of ML in offender profiling and aiding in fairer recidivism prevention strategies. The code and data are publicly available on GitHub at https://github.com/cavusmuhammed68/Recidivism-Clustering-Network-RCN-.

1. Introduction

Active crime individuals have created material and moral loss for society and the state. Recidivism, or repeated crimes, is believed to increase this material loss the most. Public spending related to material burdens caused by recidivism includes such costs as incarceration, court, and police activities. For example, it has been noted that each inmate costs the state every year in the U.S. between 30,000 and 60,000 dollars [1]. Apart from the financial burden, recidivism erodes trust in society and disrupts the very social fabric. In this respect, the research on recidivism prediction represents some of the most concrete examples of efforts to lighten both the material and moral burdens from society and governmental authority structures in terms of social cohesion and crime prevention. During the last decade, several important studies have been carried out in this field.

In this regard, the present research proposes a model referred to as the Recidivism Clustering Network (RCN), which effectively combines deep learning (DL) techniques, clustering methods, and explainable AI (XAI) in recidivism prediction studies. This effective approach is expected to provide better results with the aid of AI in criminal justice. In forecasting studies, success cannot only be estimated based on accuracy regarding the particular attributes of a given dataset. Therefore, a goal of the RCN forecasting methodology is not just an increase in accuracy but the introduction of accountability and explaining how the forecast of the results is reached. As mentioned earlier, Bharati (2024) discusses how integrating Artificial Intelligence (AI) in criminal justice systems brings about a complex interplay of efficiency and ethical concerns [2]. Issues such as equity, fairness, transparency, and the safeguarding of due process arise as AI applications, from predictive policing to sentencing algorithms, reshape the landscape. Bharati's research emphasizes the need for a balanced approach that includes human oversight, regular audits, and accountability mechanisms to mitigate risks like algorithmic bias and transparency deficiencies, which could otherwise amplify inequalities within the justice system. Equity and fairness are the other major issues arising from using AI in this field. Research focusing on juvenile delinquents has highlighted the existence of prejudices in recidivism forecasts and the adverse consequences that may occur [3]. By employing SHapley Additive exPlanations (SHAP), one of the tools within XAI, the model is designed to enhance the accountability and transparency of the decision-making prediction process, thereby mitigating the identified risks while also improving predictive accuracy. Apart from the ethical issues present in the applications of AI within criminal justice, the functionality of AI itself is criticized. [4] concentrated their work on how judges make predictions about recidivism risk with the COMPAS tool and the pitfalls of this predictive process. They note that the RCN model may constitute a possible alternative to the practice under consideration today owing to its higher degree of transparency and understandability.

Human activity is the most important point in AI implementation in predictive research on criminal justice. Hence, the collaboration between humans and algorithms may impact the decision-making of humans both in positive and negative ways [5]. Guaranteeing that this effect is positive and that the judges can make more reliable sentences requires transparent and responsible schemes, like the RCN model, to be used when researching recidivism prognoses. There is also a criticism of the predictive power of prediction studies because the data for generating forecasts is taken from case files from the past [6]. This critique highlighted the necessity for the data set to be extensive and comprehensive, encompassing textual content that pertains to various aspects of the judicial process beyond mere outcomes, as well as the requirement for the data set to be continually updated and expanded. A review of recidivism research indicates that sophisticated machine learning (ML) techniques are employed in this domain [7]. From the methods discussed, it is clear that the RCN model attains better and more consistent results compared to the other methods.

Several works related to the prediction of recidivism have been presented [8–12]. The first work, in this regard, applied logistic regression on data obtained from the Dutch Penal Execution System about prisoners convicted between 2002–2008. This research estimated the chances of recidivism based on demographic and socio-economic factors. The model proposed here did not consider any technique to deal with class imbalance. Moreover, this method has various ethical shortcomings regarding complex link establishment and does not contain effectiveness in complex inter-relationships. The findings indicated that a clear trend of recidivism does exist among juveniles; it can be said that due to probation and different rehabilitation programs, the rate of re-offending decreases [13]. Another similar study presented the comparison of recidivism prediction for prisoners who were put under electronic monitoring devices in Minnesota. The research has shortcomings: no class imbalance measure, no clustering function, and focus only on the accuracy of the prediction. It suggested that "Offenders placed on electronic monitoring were 30 percent less likely to recidivate compared with offenders not being monitored" [14]. In 1996, the study of recidivism forecasting by [15] compared neural networks against traditional statistical models. The results showed no significant difference in efficiency between the two methods used. Their shortcomings are that the dataset is imbalanced, not prepped for clustering, and its internal working model is not quite transparent. In the study by [16], the ML technique was used to predict the recidivism rates of juvenile offenders. The data pre-processing for the study used the Principal Component Analysis (PCA), and the predictive modeling applied k-nearest neighbors (k-NN) and random forest (RF) algorithms with an accuracy rate of nearly 75%.

2. Survey on deep learning tools for recidivism prediction

Predicting recidivism remains a persistent challenge within the criminal justice domain, mainly due to several data-related constraints. These include a lack of high-quality labeled data, significant class imbalance between recidivists and non-recidivists, and limitations imposed by privacy and legal frameworks on data sharing [17,18]. In recent years, numerous DL approaches have been proposed to address these challenges while maintaining fairness, transparency, and predictive accuracy.

2.1. Challenges of data scarcity in deep learning for criminal justice

One of the most critical issues is the limited availability of labeled datasets. Publicly accessible data often lack comprehensive annotations, hampers the performance and training efficacy of DL models [17]. Furthermore, the distribution of recidivism data is typically imbalanced; non-recidivists significantly outnumber recidivists, leading to biased models that inherently favor predictions toward non-reoffending individuals [18].

In addition to technical limitations, privacy and ethical concerns further restrict access to granular personal information, which is essential for effective model training [19]. Compounding this, data fragmentation across jurisdictions results in heterogeneous datasets, reducing the models' ability to generalize across legal boundaries and socio-demographic contexts [20].

2.2. Deep learning techniques to address data scarcity

To mitigate these issues, several DL strategies have emerged. Data augmentation techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs) have been widely used to generate offender profiles synthetically. These methods increase the representativeness of minority classes, helping models learn more balanced representations [17,21].

Transfer learning is another promising approach, especially with the rise of legal domain-specific models such as Bidirectional Encoder Representations from Transformers for Legal Text (BERT-Legal) and Robustly Optimized BERT Pretraining Approach (RoBERTa). These models leverage pre-trained language representations to extract meaningful patterns from legal documents, reducing the dependency on large, labeled datasets [19].

To further address the scarcity of labeled data, self-supervised learning methods such as Simple Contrastive Learning of Representations (SimCLR) allow neural networks to learn rich representations from unlabeled legal texts and crime records through contrastive techniques [22]. Meanwhile, few-shot learning and meta-learning frameworks, such as Model-Agnostic Meta-Learning (MAML), have shown promise in generalizing to new cases with minimal training samples, which is crucial for emerging or underrepresented recidivism patterns [23].

Lastly, federated learning offers a privacy-preserving mechanism for collaborative training. By enabling multiple institutions to train models locally without sharing sensitive data, this method not only addresses privacy concerns but also improves the scalability and generalizability of predictive models across jurisdictions [20].

2.3. Interpretability of visualization techniques in judicial decisions

Dimensionality reduction techniques such as PCA and t-distributed stochastic neighbor embedding (t-SNE) are increasingly employed in criminal justice analytics to support the visualization of offender profiles and recidivism risk scores. While these methods are technically valuable for identifying latent patterns in high-dimensional data, their real-world applicability in judicial contexts hinges upon how much their outputs are interpretable and actionable.

PCA, by preserving variance while reducing feature dimensionality, enables the classification of offenders into distinct groups based on factors such as criminal history, demographics, and behavioral traits. This stratification supports courts in distinguishing between low-, medium-, and high-risk offenders, thereby promoting more nuanced sentencing strategies as opposed to uniform punitive approaches [19]. Building on this, t-SNE enhances interpretability by revealing natural clusters in offender populations that may not be immediately evident through traditional statistical techniques.

The clustering insights derived from these techniques can also serve as critical tools in detecting systemic biases. For example, if t-SNE visualizations consistently show that certain racial or socioeconomic groups are overrepresented in high-risk clusters, this may point to inherent bias within the underlying data or model. Such insights enable policymakers and legal professionals to audit risk assessment frameworks, ensuring greater alignment with principles of equity and non-discrimination [20].

Moreover, visual clustering allows for the tailoring of rehabilitation and intervention programs. By identifying subgroups of offenders who share similar risk factors and behavioral patterns, correctional institutions can implement more targeted educational, vocational, or therapeutic interventions rather than relying on generalized policies [23]. This approach holds particular promise for enhancing rehabilitation outcomes and reducing reoffending rates.

Another practical application lies in the development of early-warning systems. Clusters identified through these techniques may reveal recurring recidivism patterns, enabling probation officers and parole boards to prioritize monitoring and support for individuals most at risk of reoffending. This proactive strategy enhances public safety and supports the efficient allocation of resources within the criminal justice system [22].

While PCA and t-SNE are often viewed as technical tools for data exploration, their integration into judicial workflows must be framed around actionable insights. Grounding clustering results in legal and policy considerations, and these techniques can enhance fairness, reduce bias, and inform more effective rehabilitative strategies, ultimately contributing to a more data-driven and equitable justice system.

2.4. Enhancing interpretability in deep learning models using SHAP for judicial decision-making

The integration of SHAP has emerged as a powerful technique for elucidating the importance of features in ML models, particularly in complex DL architectures. Nevertheless, interpretability remains a significant challenge—especially when communicating model outputs to judges, parole boards, policymakers, and legal professionals who may lack technical backgrounds. Several strategies must be considered to ensure the effective and ethical deployment of SHAP in recidivism prediction.

First, SHAP values should be contextualized for different stakeholder groups. For judges and parole boards, it is essential to present case-specific explanations identifying which factors — such as prior convictions, age, or employment status — contributed most significantly to an individual's risk assessment. Rather than simply reporting a numerical probability (e.g., "This individual has a 70% risk of reoffending"), the model should offer an interpretable narrative: for instance, "The model assigns high risk due to a history of three prior offenses (+0.25 SHAP contribution) and unstable housing (+0.15), while stable employment (−0.10) reduces the risk". For policymakers, SHAP values can be aggregated across large datasets to detect systemic biases and assess whether current risk assessment frameworks disproportionately affect specific populations. This insight can then inform adjustments to legislation or model design.

Second, visual representation of SHAP outputs is key to improving transparency. Tools such as feature importance bar charts and waterfall plots can clearly depict individual features' positive and negative contributions to a risk score. Additionally, force plots for individual cases can demonstrate how particular variables push a prediction toward a high-risk or low-risk classification. These visual aids make AI-driven decisions more explainable and straightforward to interrogate in legal settings.

Third, SHAP explanations can be directly aligned with core legal principles. Many jurisdictions require that judicial decisions be grounded in clear, reasoned analysis. SHAP supports this requirement by offering a transparent rationale for AI-generated predictions. For

Table 1
Comparative analysis of DL tools for recidivism prediction.

DL methods	Advantages	Limitations	Applications in recidivism prediction
SMOTE (used in RCN model)	Balances class distribution, simple to implement	Can create synthetic noise	Handling class imbalance in datasets [21]
GANs & VAEs	Generate realistic offender profiles	High computational cost, risk of bias	Synthetic dataset generation [17]
Transfer Learning (BERT-Legal, RoBERTa)	Uses pre-trained knowledge, reduces labeled data need	Requires domain-specific fine-tuning	Legal text processing, case law analysis [19]
Few-Shot Learning (MAML, ProtoNet)	Learns from very few examples	Still requires well-curated small datasets	Risk assessment on new cases [23]
Federated Learning	Enables privacy-preserving AI	Complex infrastructure needed	Cross-jurisdictional crime prediction [20]
Self-Supervised Learning (SimCLR, MoCo)	Learns features from unlabeled data	Requires large pre-training corpus	Unsupervised recidivism pattern learning [22]

example, in the context of due process, SHAP helps ensure that individuals understand how their personal data has influenced the assessment and whether potential biases are present. Similarly, in sentencing or parole decisions, SHAP provides a means for auditing risk scores to verify that marginalized groups are not unfairly penalized.

Finally, two critical components must be in place to improve trust and accountability in AI applications within the criminal justice system. First, stakeholder education is necessary. Judges, attorneys, and policymakers should receive training materials that explain SHAP-based outputs in accessible and interpretable formats. Second, human oversight should always complement algorithmic decision-making. Legal professionals must review and validate risk scores generated by DL models to ensure consistency with ethical and legal standards.

While SHAP greatly enhances the interpretability of DL models in recidivism prediction, its effectiveness in judicial decision-making depends on how well the insights are communicated and integrated into legal workflows. When supported by visualizations, case-specific narratives, and alignment with legal principles, SHAP-based models can provide transparent, equitable, and trustworthy risk assessments that enhance confidence in AI-assisted decisions.

2.5. Comparative analysis of deep learning tools

Table 1 presents a comparative analysis of DL techniques for predicting recidivism. Each method offers unique advantages in handling data limitations, improving predictive accuracy, and ensuring fairness in criminal justice applications. SMOTE effectively mitigates class imbalance, while GANs and VAEs generate synthetic data to address data scarcity. Transfer learning reduces the dependency on labeled data by leveraging pre-trained models. Few-shot learning enhances generalization with minimal examples, whereas federated learning ensures privacy-preserving AI. Lastly, self-supervised learning enables feature extraction from large-scale unlabeled datasets, aiding in more comprehensive recidivism risk assessment.

3. Legal considerations in AI-driven recidivism prediction

AI risk assessment models in the criminal justice system have significantly expanded in recent years [24]. They are mainly used to assist courts in predicting recidivism, making parole and sentencing decisions, and finding the appropriate rehabilitation method [25]. All these decisions are connected to predicting recidivism — the likelihood that an individual will re-offend [26]. Although the practice of predicting recidivism is not new, the integration of ML models into this process has become increasingly prevalent due to their ability to process large datasets and potentially enhance decision-making accuracy [27] and reduce human cognitive biases [3]. However, these ML-enabled tools like COMPAS bring significant challenges to fundamental legal principles such as individual sentencing, transparency and accountability, non-discrimination and bias mitigation, and presumption of innocence.

There are other legal challenges AI poses in criminal legal proceedings, and it requires a book-scale work to discuss all of them. The legal challenges that will be discussed below were selected based on two criteria: First, their relevance to the recidivism process, meaning that they mostly affected legal principles by the use of AI in the recidivism process; and second, the extent to which the AI model presented in this paper contributes to compliance with these legal principles.

1. Principle of individualized sentencing: Individualized sentencing requires judicial discretion that respects each defendant's unique circumstances rather than relying solely on generalized trends, particularly in capital cases [33–35]. However, the "individualized" sentencing principle fundamentally contradicts the working principles of AI because what makes AI an appealing tool is its ability to make predictions based on patterns and draw "generalizations" from former cases [36]. This conflict was also highlighted in the landmark case State v. Loomis, where the Wisconsin Supreme Court ruled that AI risk scores should not be the sole determinant in sentencing and that sentencing must integrate personalized assessments beyond algorithmic outputs to preserve the individualized sentencing [37,38].
   In the face of this contradiction, two options exist to uphold the individualized sentencing principle. The first option is to abandon the use of AI altogether in order to strictly adhere to individualized sentencing, which will lead to sacrificing enormous advantages provided by AI [39]. The second option is to use an AI model capable of assessing the relevance of the several circumstances of the past cases to the individual case at hand in the most accurate way, thus minimizing any divergence from the principle.
   As illustrated in Table 2, the RCN is the AI model capable of clustering numerous inputs by its technique. With this capacity, it assesses the circumstances of the individual cases more accurately by putting them into the "right" cluster. Rather than overgeneralizing from past data, the RCN allows for more tailored predictions for each defendant. In that way, it bridges the gap between general data patterns and the unique characteristics of individual cases. It would enable courts and judges to receive support from AI while still upholding compliance with individualized sentencing to a large extent.
2. Principle of transparency and accountability: Transparency is another core tenet of due process [37,38]. It requires that state actions — especially sentencing — are rooted in publicly accessible, clear, and accountable procedures. For AI risk assessment tools, this transparency involves revealing how risk scores are derived, including any limitations or biases in the tool, and ensuring that defendants and legal professionals can interpret and, if necessary, challenge AI-based findings [40,41]. In Gardner v. Florida, the Court emphasized that defendants must be able to understand how AI-informed scores influence their sentencing [42].

Table 2
Comparison of our proposed method with traditional methods.

Method	Key Features	Interpretability	Ethical/Transparency Considerations	Ref.
RCN	Combines DL with clustering (k-means, PCA, t-SNE), SHAP for explainability	High – SHAP insights on feature importance	High – Transparent, addresses bias	Our Paper
Logistic Regression (LR)	Uses static predictors for 2-year predictions	Low – Limited by linear assumptions	Moderate – Transparent but limited	[13]
RF and Decision Tree Models	Uses historical data, handles non-linear relationships	Moderate – Limited feature importance analysis	Moderate – Can overfit, limited transparency	[14,28]
Neural Networks	Models complex interactions	Low – Black-box, hard to explain	Low – Hard to interpret, ethics concerns	[15]
COMPAS	Widely used for risk assessment	Low – Black-box, lacks feature transparency	Low – Criticized for bias, transparency issues	[7,29]
SHAP-enhanced Models	Gradient Boosting, RF with SHAP	High – SHAP for feature impact	High – Transparent with SHAP, limited by feature engineering	[30,31]
RisCanvi (ML-enhanced)	Predicts violent recidivism	Moderate – Lacks interpretability tools	Moderate – Potential for bias	[5,32]
RF for Youth Offenders	Uses administrative data for youth recidivism	Moderate – Limited transparency	Moderate – Limited fairness insights	[16]

Legal Basis for SHAP in Judicial Decision-Making

The inclusion of SHAP in the RCN framework directly addresses judicial concerns surrounding explainability, transparency, and accountability — core principles in both U.S. constitutional due process and international human rights standards. Several landmark legal cases and policy frameworks support using interpretable AI systems, including SHAP-enhanced models in high-stakes domains such as criminal justice.

First, the U.S. Supreme Court in *Gardner v. Florida*, 430 U.S. 349 (1977), held that defendants have a right to be informed about the basis of sentencing decisions, particularly when those decisions are influenced by undisclosed data or assessments. SHAP values provide precise, individualized explanations for risk assessments, aligning with this precedent by making model predictions contestable and transparent [42].

In *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), the Wisconsin Supreme Court warned against the opaque nature of the COMPAS tool, noting that defendants could not meaningfully challenge how specific inputs (such as age or gender) influenced their risk scores. Although the court allowed COMPAS, it emphasized the need for caution and transparency. SHAP helps fulfill this legal expectation by quantifying the contribution of each feature in a clear and explainable way [43].

International policy frameworks echo this demand for explainable AI:

- The *EU Guidelines for Trustworthy AI* stress the importance of transparency, explainability, and accountability, particularly for high-risk applications such as criminal justice [44].
- The *OECD Principles on AI* advocate for "transparency and responsible disclosure" in AI systems to ensure meaningful human oversight [45].
- The *White House Blueprint for an AI Bill of Rights* (2022) identifies "notice and explanation" as a foundational principle, asserting that individuals must know when an automated system is in use and receive clear explanations of outcomes [46].

Legal scholars have also emphasized the critical need for explainable systems in justice applications. Citron and Pasquale argue that "black-box" scoring systems violate core due process protections when they are inscrutable to affected individuals [47]. SHAP directly addresses this issue by offering human-readable, feature-level attribution of decisions. In light of these legal and ethical imperatives, SHAP is not only a technical tool but also a legal safeguard. It ensures that algorithmic risk assessments — like those performed by the RCN — are intelligible, justifiable, and contestable, reinforcing the integrity of judicial decision-making processes.

The RCN addresses transparency through the integration of XAI techniques including SHAP, which allows judges, defendants, and legal professionals to ascertain how specific factors contribute to the risk assessment, unlike COMPAS, which operates as a "black box" model — meaning that they are too opaque for those who do not have advanced technical knowledge [48]. The RCN's use of SHAP provides a detailed breakdown of how various inputs contribute to an individual's score. This transparency allows defendants to contest or question the fairness of their assessment, which upholds the right to due process. Additionally, the RCN's transparent design also aids legal professionals in making well-informed sentencing recommendations, thereby enhancing the legitimacy and fairness of AI-assisted decisions [37,38].

3. Principle of non-discrimination and bias mitigation: A critical requirement for AI in the criminal justice system is adherence to the principle of non-discrimination by avoiding any potential for bias [49,50]. In the State v. Loomis case, Loomis, the defendant, argued the unlawful inclusion of certain variables such as gender, age, or race as inputs by the COMPAS [37,38]. However, this has not been considered an acceptable objection by the Wisconsin Supreme Court as such a move would have a negative effect on the accuracy of the outcomes produced by AI systems [37,38]. Furthermore, even when explicit variables such as gender or race are removed, AI models can exhibit "proxy discrimination" by associating other seemingly neutral variables with these sensitive attributes [51].

   The RCN minimizes the bias risks through its advanced clustering techniques. Using SHAP to evaluate the weight of each input, the RCN can ensure that predictions are not biased. This technique also allows the model to identify latent patterns within the data while remaining sensitive to the risk of proxy discrimination. For instance, the RCN's clustering method can recognize and factor in patterns related to individual behavior, education level, or employment status without relying on demographic proxies that might perpetuate biases. By focusing on personal characteristics and behavioral data rather than static demographic factors, the RCN can provide a fairer assessment of each individual's likelihood of recidivism.
4. Principle of presumption of innocence and accuracy: The presumption of innocence is another principle closely related to

using the AI recidivism score as evidence in criminal sentencing processes. According to the presumption of innocence, the onus is on the court to prove that the accused is guilty of certain crime(s) [52]. As reflected in the words of Sir William Blackstone, the presumption of innocence envisages that "it is better that ten guilty persons escape than that one innocent suffer" [53]. However, many AI models designed for recidivism prediction, including COMPAS and HART, are often designed to err on the side of caution [39]. In other words, they are more likely to flag low-risk individuals as high-risk, subjecting them to unnecessary punitive measures. This misclassification can result in excessive punitive measures and contradict the presumption of innocence and fairness the legal system aims to uphold.

In response to that challenge, the RCN achieves almost 75% accuracy, surpassing the predictive reliability of previous models while avoiding ethical compromises. By employing clustering techniques that analyze individual behavioral data and using SMOTE to address class imbalances, the RCN maintains high prediction accuracy without sacrificing fairness. The model's ability to balance precision and ethical responsibility helps to preserve the presumption of innocence. Rather than over-relying on demographic predictors, the RCN's emphasis on behavioral data allows for more accurate, ethically sound predictions and reduces the risk of unjustly categorizing defendants. This approach is compatible with the fundamental commitment of the justice system to protecting the innocent.

The challenges surrounding AI-based recidivism models, such as lack of transparency, risk of discrimination, and erosion of individualized sentencing, underscore the need for a more legally compliant AI model. The RCN seeks to address these issues by establishing a framework that aligns closely with essential legal principles while utilizing the efficiencies of AI.

First, the RCN advances compliance with the principle of individualized sentencing by refining the clustering of individual factors. This enables incorporating more diverse circumstances into risk assessment while preserving the balance between judicial discretion and predictive analytics. Second, the RCN enhances transparency. Unlike existing models, the RCN offers transparency through SHAP, which allows all parties involved in sentencing to understand and, when necessary, challenge the AI's risk assessment. Third, the RCN mitigates the bias. Increasing the accuracy of its advanced clustering technique provides less biased demographic information. In this way, it minimizes the risk of both direct and indirect biases in assessments, supporting non-discriminatory outcomes. Last but not least, it provides a fairer result. By achieving about 75% accuracy, the RCN offers an AI model that ensures better compliance with the presumption of innocence, reducing false positives and avoiding unnecessary punitive measures for low-risk individuals.

The RCN represents a significant advancement in the ethical integration of AI-based recidivism assessments within the criminal justice system. By prioritizing the principles of individualized sentencing, transparency, non-discrimination, and the presumption of innocence, RCN addresses longstanding challenges associated with traditional models such as COMPAS. The model's innovative use of explainability and advanced clustering techniques enhances judicial decision-making by aligning AI predictions with foundational legal standards. This approach fosters greater trust and fairness within the justice process. Therefore, RCN sets a new benchmark for AI-based risk assessment by providing an accountable, accurate, and ethically sound tool that empowers the justice system to make informed, responsible sentencing decisions.

3.1. Key benefits of the RCN method compared to traditional models

As shown in Table 2, the RCN's DL and clustering approach captures more complex, non-linear data relationships than LR's limited, linear model. RCN's ability to uncover latent clusters provides deeper insights into offender subgroups, which LR does not. Additionally, the RCN incorporates SMOTE to handle class imbalance, improving recidivism prediction, whereas LR struggles with imbalanced datasets [13]. While RF models handle non-linear relationships well, RCN outperforms in terms of explainability and interpretability by integrating SHAP values, which provides clearer insights into why predictions are made. RF models often suffer from interpretability issues and can be prone to overfitting, while RCN addresses this with its clustering approach and XAI tools [14,28]. The RCN's combination of DL with clustering methods provides better explainability and insights than traditional neural networks, which are often viewed as black-box models. RCN is more interpretable due to the integration of SHAP values, and it leverages clustering techniques along with k-means, PCA, and t-SNE to reveal latent patterns hidden in purely neural network-based models [15]. COMPAS is widely criticized for bias and lack of transparency, whereas RCN addresses these issues with XAI (via SHAP) and more interpretable predictions. While COMPAS is a powerful tool, it lacks the transparency and fairness mechanisms RCN provides, making RCN a more ethical and transparent choice for criminal justice systems [7,29]. RCN extends the advantage of SHAP by incorporating clustering methods, which these SHAP-based models do not. While both RCN and SHAP-enhanced models provide transparency, RCN offers the additional ability to cluster individuals into subgroups, providing deeper insights into different offender profiles that these models do not [28,31]. While RisCanvi uses ML to predict recidivism, RCN surpasses it with its DL and clustering methods, providing more nuanced insights into the offender population. RCN's ability to handle data imbalance with SMOTE and its use of clustering techniques enhances prediction quality, while RisCanvi focuses more on logistic regression and lacks such advanced techniques [32]. Similar to general RF models, the RCN model outperforms in terms of interpretability and clustering capabilities. RCN is also more adept at handling complex data relationships and provides clear explanations for predictions, while RF models applied to youth offenders do not offer as much insight into the underlying data patterns [16].

3.2. Impactful contributions of this research in AI-based recidivism analysis

This paper makes several key contributions to recidivism prediction and the broader intersection of AI, criminology, and predictive analytics. The main contributions of this study are as follows:

1. Introduction of the RCN: We propose a RCN framework which integrates DL and clustering techniques. The RCN method is distinct from traditional models by employing clustering algorithms, such as k-means, in combination with PCA and t-SNE to uncover latent patterns in offender populations. This layered architecture allows for more accurate predictions of recidivism while also offering insight into the subgroups within the offender population that are at higher risk of re-offending.
2. Addressing data imbalance with SMOTE: The model tackles the significant challenge of data imbalance, which is prevalent in recidivism datasets where non-recidivists far outnumber recidivists. By incorporating the SMOTE, the RCN ensures fairer and more balanced predictions. This technique mitigates the bias seen in many ML models, which typically favor the majority class, by oversampling the minority class, thus improving model performance for recidivists.
3. Integration of XAI techniques: The study integrates SHAP into the RCN framework, making the black-box nature of DL models more interpretable. This addresses the ethical concerns surrounding AI models in criminal justice systems, where transparency is crucial. SHAP values provide insights into which features (e.g., age, prior convictions) most influence recidivism predictions, ensuring that decisions made by the model can be scrutinized and explained to stakeholders.

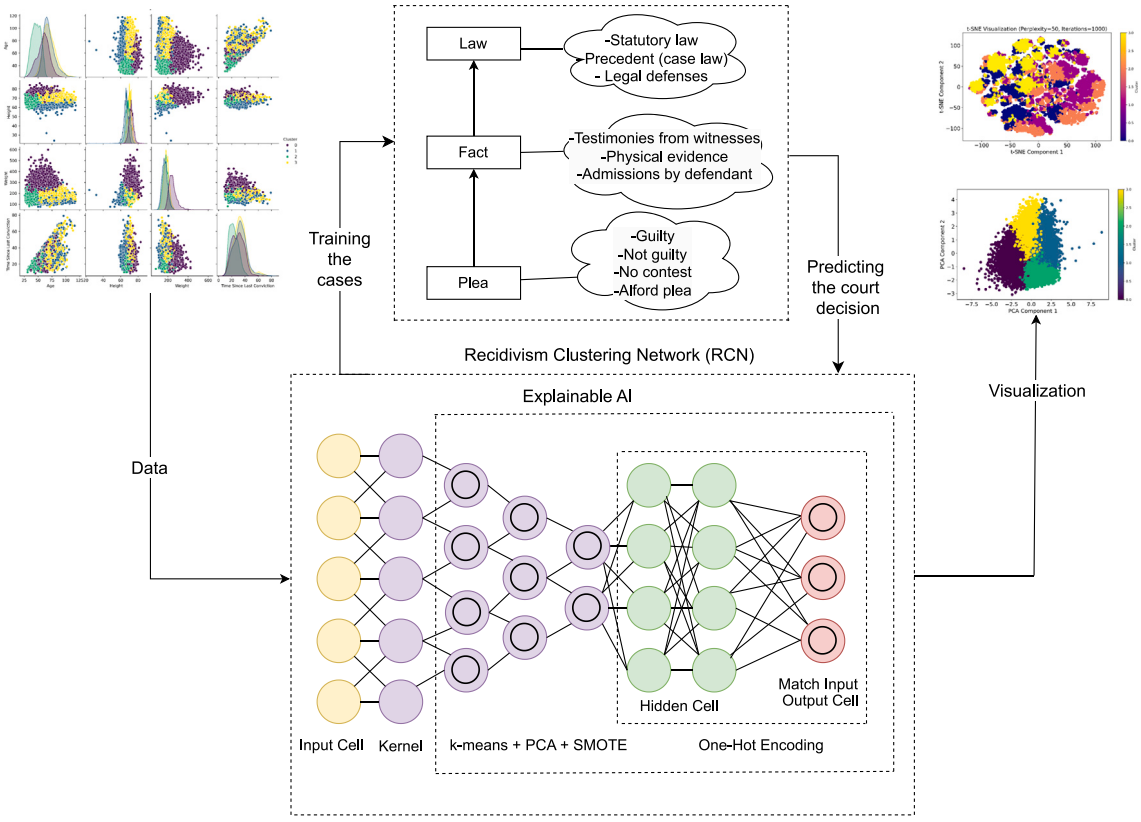


Fig. 1. The RCN framework: Integrating ML and legal decision-making.

Fig. 1 presents an integrated framework for predicting recidivism using the RCN, which combines legal data and ML techniques. The process begins with data preprocessing, where essential features such as age, height, and categorical variables are converted into numerical formats. This preprocessing step ensures that the data is consistent and ready for analysis. The model handles any missing data through imputation and scales the numerical features to improve performance in later model stages. The data then goes through k-means clustering, PCA, and SMOTE for data balancing and dimensionality reduction, ensuring optimal performance on balanced and relevant feature sets. The RCN is the core of the system, where input data is processed through multiple layers, including input cells, hidden layers (kernel cells), and output cells. The network learns complex relationships between the features, identifying patterns associated with recidivism. Through One-Hot Encoding, categorical variables such as gender and ethnicity are transformed into a format the neural network can understand, enhancing the model's ability to interpret a wide range of inputs. A key aspect of this framework is the integration of XAI. This component ensures that the RCN's decision-making process remains transparent, allowing legal professionals to interpret and understand the factors influencing the model's predictions. The transparency XAI provides is crucial in legal contexts, where model-driven decisions must be justifiable and interpretable. Additionally, the model is trained on actual legal cases, incorporating important legal inputs such as pleas (e.g., guilty, not guilty), facts (e.g., witness testimony, physical evidence), and laws (e.g., statutes, precedents, legal defenses). This legal context enhances the model's relevance to real-world court decisions. After training, the model can predict outcomes such as the likelihood of recidivism based on prior convictions, release history, and demographic data. The final output includes both model predictions and visualizations, such as t-SNE and PCA, which provide a two-dimensional representation of the data clusters, offering a clear visual interpretation of offender patterns. These visual outputs help understand the relationships between different offender attributes and recidivism risks. Legal professionals can make more informed decisions through the RCN, leveraging data-driven insights to assess recidivism risks while ensuring the model remains interpretable and trustworthy.

3.3. Practical integration of RCN in judicial workflows

The RCN framework is designed for academic research and practical deployment within legal and judicial systems. RCN can be integrated into existing risk assessment processes, including tools such as COMPAS, to support fairer and more transparent sentencing decisions. The following points outline how RCN could be operationalized in real-world settings:

- **Data Interface Layer:** RCN can process the same administrative and legal datasets typically used in current judicial workflows. These include demographic details, criminal history, and contextual legal information. Through application programming interfaces (APIs), RCN can connect with existing court management software or parole board systems to enable seamless data integration.
- **Parallel Assessment with COMPAS:** RCN can complement existing models, such as COMPAS. Risk scores generated by both systems can be compared in real time. The key advantage of RCN lies in its use of SHAP-based explainability, which offers clear, interpretable insights into how each feature contributes to a risk score — unlike COMPAS, which is often criticized for being a black-box model.
- **Explainable AI for Legal Stakeholders:** One of RCN's strengths is its ability to generate case-specific explanations. RCN supports judges, parole boards, and attorneys in understanding the reasoning behind predictions by quantifying how features such as the number of prior convictions or time since the last release influence the risk score. This fosters trust and aligns with the principles of due process.

Table 3
Descriptive statistics of offender characteristics and criminal history.

	Year of last conviction	Year of last release	Height	Weight	Age
Count	3920	3920	49 446	49 446	49 446
Mean	1994.14	1996.62	69.10	191.63	61.84
Std	9.94	10.13	3.27	42.35	14.12
Min	1945	1949	24	75	26
25%	1988	1990	67	162	52
50%	1994	1997	69	185	62
75%	2002	2005	71	212	71
Max	2015	2100	86	600	118

- **Clustering-Based Offender Profiling:** RCN employs unsupervised learning techniques, including k-means clustering, PCA, and t-SNE, to identify subgroups of offenders with similar characteristics. These clusters can inform the development of tailored rehabilitation programs, targeted supervision strategies, and risk-informed judicial decisions.
- **Policy and Reform Applications:** Beyond individual assessments, RCN outputs can identify systemic issues, such as disproportionately representing certain demographic groups in high-risk categories. This enables data-driven policy reforms to improve fairness and equity in criminal justice practices.

RCN bridges the gap between advanced predictive modeling and foundational legal principles by combining deep learning, clustering, and explainability in a transparent framework. Its integration into the existing legal infrastructure can enhance both the accuracy and ethical integrity of AI-assisted decision-making in the justice system.

4. Developing the recidivism clustering network

4.1. Detailed overview of dataset and demographics

The dataset sourced from the California Megan's Law website contains records on 49,446 individuals, each detailing a range of personal and legal information. Key fields include a description of offenses and corresponding offense codes, although some records lack risk assessment scores. Demographic details such as date of birth, ethnicity, eye color, hair color, height, and weight are provided for each individual. Additionally, records show the last known address and sex and may include specific administrative notes or warnings, such as registration violations. This comprehensive information aims to support law enforcement and enhance public awareness regarding registered offenders, though some data fields (e.g., risk scores) are incomplete [54]. The descriptive statistics in Table 3 provide key insights into the characteristics of offenders in terms of age, height, weight, and criminal history, specifically focusing on the "Year of last conviction" and "Year of last release". The data [54] is based on 49,446 individuals for physical attributes and 3,920 individuals for conviction-related variables. The average age of the offenders is approximately 61.8 years, with a standard deviation of 14.1 years, indicating a wide age distribution, with some individuals as old as 118 years. Heights range from 24 to 86 inches, with a mean of 69.1 inches (about 5 ft 9 inches), while the average weight is around 191.6 pounds, with a significant range from 75 to 600 pounds. Conviction data shows that the average year of the last conviction is 1994, and the average year of the last release is 1996, with a few outliers showing years as recent as 2100, which may indicate incorrect data or outliers. The standard deviations for these years, around 9.9 and 10.1, respectively, suggest some variability in the timing of these events. The interquartile ranges (25th to 75th percentiles) indicate that most offenders were convicted between 1988 and 2002 and released between 1990 and 2005. These statistics provide a broad overview of the dataset, highlighting not only the physical characteristics of offenders but also the temporal distribution of their criminal history.

Fig. 2 illustrates a detailed breakdown of offender ages, with the majority falling between 40 and 80 years—the distribution peaks around the age of 60, where approximately 7,000 offenders are recorded. The data shows a right-skewed distribution, meaning older individuals are more prevalent in this dataset. Offenders in their 50 s and 60 s dominate, as the count significantly rises after age 40 and remains high until the mid-70s. A sharp decline is observed beyond age 80, with very few offenders recorded past age 90. The minimum age in the dataset is around 20, while the oldest offender is 118 years old, but these extreme values represent outliers. This indicates that recidivism or criminal activity is more frequent among middle-aged and older individuals in this dataset. The correlation heatmap in Fig. 2 reveals the relationships between several variables in the dataset, including age, height, weight, year of last conviction, and year of last release. The most notable correlations are:

- Year of last conviction and year of last release: These two variables are highly correlated with a value of 0.94, which is expected since the year of release typically follows the year of conviction.
- Age and year of last conviction/release: Age shows a significant negative correlation with both the year of last conviction (−0.54) and the year of last release (−0.52), indicating that older individuals tend to have had earlier convictions and releases.
- Height and weight: A moderate positive correlation (0.45) between height and weight suggests that taller offenders tend to weigh more.
- Other relationships: The other variables, such as the correlations between age, height, and weight, show weak or negligible correlations, indicating limited direct relationships among these attributes.

This heatmap helps identify significant patterns and relationships within the dataset, particularly the strong temporal correlations and demographic factors.

4.2. Data analysis — feature distributions across recidivism classes

The dataset contains records on 49,446 individuals and is divided into training and validation sets to ensure proper model evaluation. The splitting is performed as follows:

- **Training Set:** 80% of the dataset is used for training the model. This ensures sufficient data for learning patterns and relationships.
- **Validation Set:** 20% of the dataset is set aside for validation, allowing us to evaluate the model's performance on unseen data.

This split is implemented using stratified sampling to maintain the proportion of recidivists and non-recidivists in both subsets, ensuring the model generalizes well across classes. Mathematically, let D represent the full dataset, then:

$$D_{\text{train}} = 0.8 \cdot D, \quad D_{\text{val}} = 0.2 \cdot D \tag{1}$$

This ensures that the training dataset contains approximately 39,557 records, and the validation dataset includes about 9,889

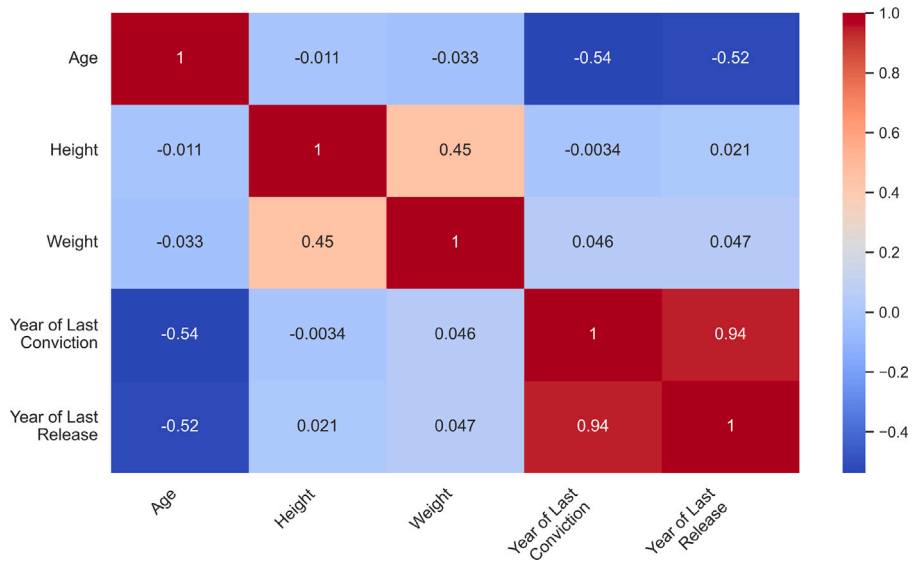


Fig. 2. Correlation heatmap of age, height, weight, year of last conviction, and year of last release.

records. Stratified sampling ensures that the class imbalance in the dataset does not disproportionately affect either subset.

Fig. 2 presents a histogram of offender ages to provide insights into the age distribution across the dataset. The following key observations can be made:

- **Age Range:** Most offenders fall between 40 and 80 years of age, peaking at around 60 years.
- **Skewness:** The distribution is right-skewed, indicating a higher prevalence of older individuals in the dataset.
- **Outliers:** A few extreme values are observed, including individuals over 100 years old, which may represent data entry errors or rare cases.

This histogram helps illustrate the demographic concentration of offenders in the dataset and informs feature engineering for models. Understanding the age distribution is particularly useful for identifying patterns of recidivism associated with different age groups. To start, we examined the distribution of key offender characteristics, such as age, height, and time since the last conviction, across recidivism classes (0 for non-recidivists and 1 for recidivists). These distributions provide insights into the relationships between these variables and recidivism outcomes.

- Age distribution: Fig. 3(a) shows the age distribution for recidivists and non-recidivists. The median age for both groups is close to 60 years. However, recidivists exhibit a slightly narrower range, with the interquartile range (IQR) spanning approximately 50 to 70 years, while non-recidivists demonstrate a slightly wider IQR, extending from around 50 to 80 years. Outliers are more prominent within the non-recidivist group, with several individuals over 100 years old. Although the median ages are relatively comparable, the distribution suggests that recidivists tend to be somewhat younger, particularly within the upper quartiles. This modest difference could reflect age-related factors in recidivism likelihood.
- Height distribution: The height distributions for recidivists and non-recidivists, shown in Fig. 3(b), are nearly identical. The median height for both groups is about 70 inches. The IQR for both recidivists and non-recidivists spans roughly 65 to 72 inches, indicating that most offenders fall within this height range, regardless of recidivism status. There are a few outliers with heights below 50 inches, particularly among non-recidivists, but the overall pattern suggests that height is not a distinguishing factor between recidivists and non-recidivists.
- Time since last conviction distribution: Fig. 3(c) illustrates the distribution of time since the last conviction for both recidivists and non-recidivists. The median time since the last conviction is around 30 years for both groups. However, the IQR for non-recidivists is slightly broader, ranging from about 20 to 40 years, compared to recidivists, where the IQR spans from 25 to 40 years. Additionally, the non-recidivist group includes more extreme outliers, with some individuals reporting over 60 years since their last conviction. These findings suggest that, while the time elapsed since the last conviction is similar between recidivists and non-recidivists, the broader distribution in non-recidivists may indicate that individuals with very long periods since their previous offense are less likely to re-offend.

4.3. Comprehensive mathematical model for the RCN method

The RCN integrates clustering techniques with DL networks to enhance recidivism prediction. The methodology consists of multiple stages: data preprocessing, clustering, neural network prediction, and model optimization. Each step is mathematically formalized, with equations linking various processes.

4.3.1. Data preprocessing and feature engineering

Given a dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ containing n samples (offenders) and d features, preprocessing is performed to prepare the data for clustering and prediction.

Numerical feature transformation: Let $\boldsymbol{x}_i$ represent the numerical feature vector for offender i. Each feature is standardized by subtracting the mean $\mu(\boldsymbol{x})$ and dividing by the standard deviation $\sigma(\boldsymbol{x})$. The transformed feature vector $\boldsymbol{x}'_i$ is:

$$\boldsymbol{x}'_i = \frac{\boldsymbol{x}_i - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \tag{2}$$

Categorical feature transformation: For categorical variables, one-hot encoding is applied. Let c_j represent a categorical feature with k categories. The one-hot encoded feature vector $\text{OneHot}(c_j)$ is:

$$\text{OneHot}(c_j) = [0, \ldots, 1, \ldots, 0] \tag{3}$$

The final pre-processed feature vector $\boldsymbol{x}'_i$ for each offender combines the numerical and one-hot encoded categorical features:

$$\boldsymbol{x}'_i = [\boldsymbol{x}_{\text{numerical}}, \boldsymbol{x}_{\text{categorical}}] \tag{4}$$

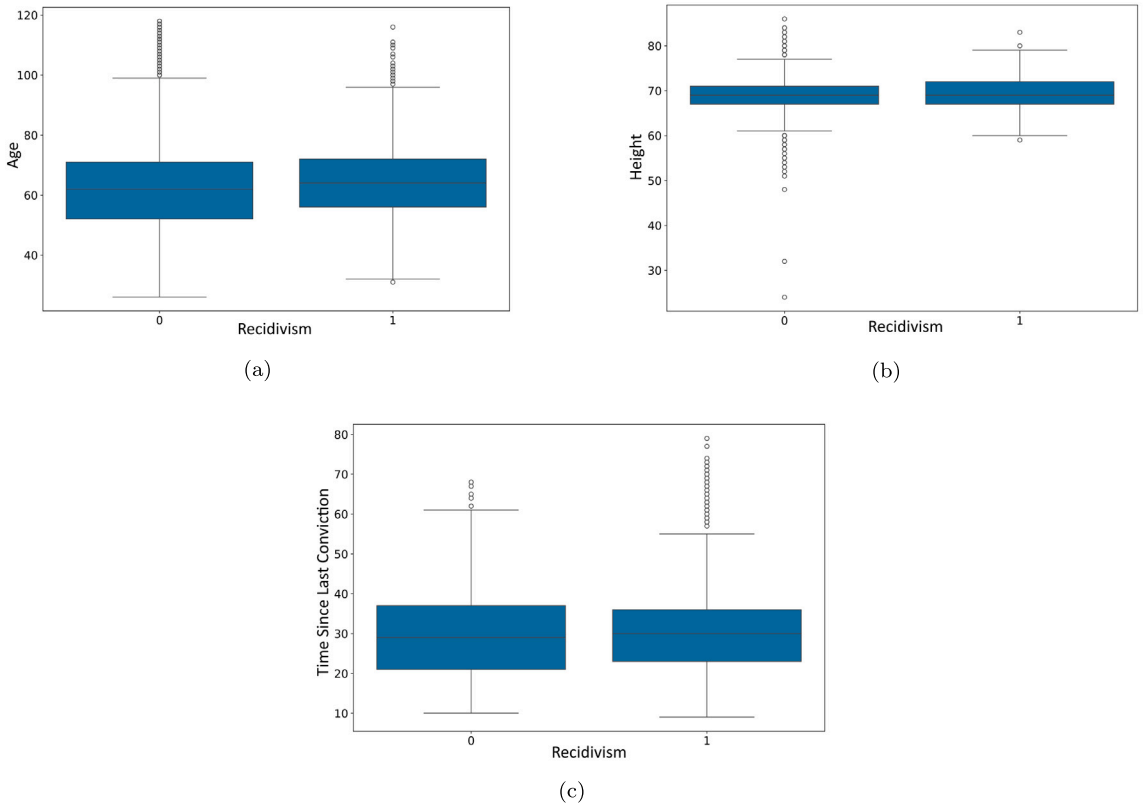


Fig. 3. Distributions of various characteristics across recidivism classes. (a) Age distribution, (b) Height distribution, (c) Time since last conviction distribution.

4.3.2. Clustering layer

Clustering is performed on the pre-processed feature vectors $\boldsymbol{x}'_i$. k-means clustering divides the data into K clusters by minimizing the within-cluster variance:

$$\min_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_K} \sum_{i=1}^{n} \min_{k\in\{1,2,\ldots,K\}} \|\boldsymbol{x}'_i - \boldsymbol{\mu}_k\|^2 \tag{5}$$

where $\boldsymbol{\mu}_k$ is the centroid of the kth cluster, and K is the total number of clusters. The clustering results in partitioning the dataset into clusters $C = \{C_1, C_2, \ldots, C_K\}$, where each cluster C_k contains a subset of the offenders.

4.3.3. Recidivism prediction with neural networks

After clustering, a deep neural network is trained to predict recidivism for offenders in each cluster. Each cluster C_k is associated with its neural network, parameterized by weights $\boldsymbol{\theta}_k$.

Neural network architecture: The neural network for cluster C_k consists of L layers. Let $f_{\boldsymbol{\theta}_k}(\boldsymbol{x}'_i)$ denote the output of the neural network for offender i in cluster C_k. The transformation through layer l is given by:

$$\boldsymbol{z}_l = \text{ReLU}(\boldsymbol{W}_l \boldsymbol{z}_{l-1} + \boldsymbol{b}_l) \tag{6}$$

where $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ are the weight and bias matrices of layer l, respectively, and $\boldsymbol{z}_0 = \boldsymbol{x}'_i$. The ReLU activation function is defined as $\text{ReLU}(x) = \max(0, x)$.

The final layer produces the predicted probability of recidivism $\hat{y}_i$ for offender i:

$$\hat{y}_i = \sigma(\boldsymbol{W}_L \boldsymbol{z}_{L-1} + \boldsymbol{b}_L) \tag{7}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function.

SHapley Additive exPlanations: SHAP enhances the interpretability and transparency of neural network predictions. SHAP values are derived from cooperative game theory, and the contribution of each feature is allocated fairly to the model's predictions. The SHAP value for a feature i is defined as:

$$\phi_i = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}\left[f(S\cup\{i\}) - f(S)\right] \tag{8}$$

where N is the set of all features, S is a subset of features excluding the feature i, and $f(S)$ is the model's output when only the features in subset S are included.

This formula ensures that the contribution of each feature is fairly distributed across all possible subsets of features.

In the RCN, SHAP values are computed to analyze how each feature, such as *age*, *prior convictions*, and *time since last conviction*, contributes to the predicted likelihood of recidivism. These explanations allow stakeholders, such as legal professionals, to understand and trust the model's predictions.

Integration into the RCN Framework: In the RCN, SHAP values are calculated for each prediction to ensure:

1. Transparency: Providing insights into which features influence the prediction.
2. Bias Mitigation: Ensuring the model's decisions are interpretable and not driven by proxies for sensitive attributes.
3. Actionability: Enabling legal stakeholders to contest or validate the prediction based on the contributions of specific features.

By incorporating SHAP values into the RCN, the model improves interpretability and adheres to the principles of transparency and accountability essential in criminal justice applications.

Example Calculation of SHAP: Consider a simplified dataset with three features: $X = \{x_1, x_2, x_3\}$. For feature x_1, its SHAP value ϕ_1 would involve evaluating the marginal contribution of x_1 for all possible subsets of the remaining features $\{x_2, x_3\}$:

$$\phi_1 = \frac{1}{3}\left[f(\{x_1,x_2,x_3\}) - f(\{x_2,x_3\})\right] + \frac{1}{3}\left[f(\{x_1,x_2\}) - f(\{x_2\})\right] + \frac{1}{3}\left[f(\{x_1,x_3\}) - f(\{x_3\})\right]$$

Table 4
Comparison of neural network and RF performance.

Metric	Neural network	Random forest
Accuracy (%)	75	73
Precision (%)	72	70
Recall (%)	78	75
F1-score (%)	75	72
Feature Importance Analysis	Limited (via SHAP)	Available (natively)
Interpretability	Moderate (via XAI)	High

Here, the weights for each subset are calculated based on the size of the subset, as shown in Eq. (8).

Random Forest (RF) Modeling in the RCN Framework: In addition to neural networks, RF is incorporated as a baseline model in the RCN framework. RF is an ensemble learning method that builds multiple decision trees and combines their outputs to make predictions. This ensures robustness to overfitting and improves accuracy, especially in datasets with a complex feature space like the recidivism dataset.

The RF prediction for recidivism, $\hat{y}$, is computed by averaging the outputs of B decision trees:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(X) \tag{9}$$

where B is the total number of trees and $T_b(X)$ is the prediction from the bth decision tree given input features X.

RF is particularly effective in handling high-dimensional datasets and provides feature importance metrics, allowing for the interpretability of the factors driving recidivism predictions. For example, age and prior convictions emerged as dominant features in the RF model, aligning with SHAP insights from the neural network.

Comparison between neural networks and random forest

The RCN framework evaluates both neural networks and RF for recidivism prediction to assess trade-offs between predictive performance and interpretability. Table 4 summarizes their performance on key metrics:

RF is an auxiliary model in the RCN workflow, complementing neural networks in identifying key features driving recidivism. While the neural network excels in modeling complex non-linear relationships, RF is leveraged to cross-validate feature importance and ensure robustness in predictions.

The combination of RF and neural networks provides a comprehensive framework for recidivism prediction:

1. RF offers interpretable feature importance rankings and baseline predictions.
2. Neural networks, with their clustering and residual learning layers, provide deeper insights into latent data patterns.

The study balances explainability and advanced predictive capabilities by incorporating RF into the RCN framework.

4.3.4. Loss function and optimization

The model is trained to minimize the binary cross-entropy loss function. For each cluster C_k, the loss function $L(\boldsymbol{\theta}_k)$ is:

$$L(\boldsymbol{\theta}_k) = -\frac{1}{N_k}\sum_{i\in C_k}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{10}$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N_k is the number of offenders in the cluster C_k.

Gradient descent update rule: The model parameters $\boldsymbol{\theta}_k = \{\boldsymbol{W}_l, \boldsymbol{b}_l\}_{l=1}^{L}$ are updated using stochastic gradient descent (SGD). The parameter update at iteration t is given by:

$$\boldsymbol{\theta}_k^{(t+1)} = \boldsymbol{\theta}_k^{(t)} - \eta \nabla_{\boldsymbol{\theta}_k} L(\boldsymbol{\theta}_k) \tag{11}$$

where η is the learning rate and $\nabla_{\boldsymbol{\theta}_k} L(\boldsymbol{\theta}_k)$ is the gradient of the loss with respect to the parameters. Combining the clustering formulation in Eq. (5) with the neural network prediction process described in Eqs. (6) and (7), the RCN methodology tailors the recidivism prediction models to specific offender groups.

4.3.5. Residual learning in RCN

To improve the training of deep networks, residual learning is incorporated. The residual block allows the input $\boldsymbol{x}_i'$ to bypass layers, ensuring that deeper layers can focus on learning residual mappings. The residual learning equation is:

$$y_i = F(\boldsymbol{x}_i') + \boldsymbol{x}_i' \tag{12}$$

where $F(\boldsymbol{x}_i')$ represents the transformation through the hidden layers and $\boldsymbol{x}_i'$ is the input to the residual block. This addition enables the model to learn complex representations while mitigating the vanishing gradient problem.

4.3.6. Final model output

The output of the cluster-specific neural network gives the final predicted probability of recidivism for each offender i:

$$\hat{y}_i = f_{\boldsymbol{\theta}_{C(\boldsymbol{x}_i')}}(\boldsymbol{x}_i') \tag{13}$$

where $C(\boldsymbol{x}_i')$ represents the cluster assignment for offender i, and $f_{\boldsymbol{\theta}_{C(\boldsymbol{x}_i')}}(\boldsymbol{x}_i')$ is the neural network associated with cluster $C(\boldsymbol{x}_i')$.

4.3.7. Evaluation metrics

The performance of the RCN is evaluated using several metrics:

Accuracy:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{14}$$

Precision and Recall:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{16}$$

This measure indicates the model's ability to identify all relevant cases correctly. High recall ensures that most of the positive cases are captured, even at the cost of some false positives.

F1-score: The harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{17}$$

ROC-AUC: The area under the ROC curve, plotting the true positive rate (TPR) against the false positive rate (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}\, d(\text{FPR}) \tag{18}$$

These equations formalize the methodology of the RCN and demonstrate how clustering and DL are combined to enhance predictive performance. The RCN method addresses heterogeneity in the data and improves recidivism prediction by assigning each offender to a specific cluster and tailoring the neural network to that cluster.

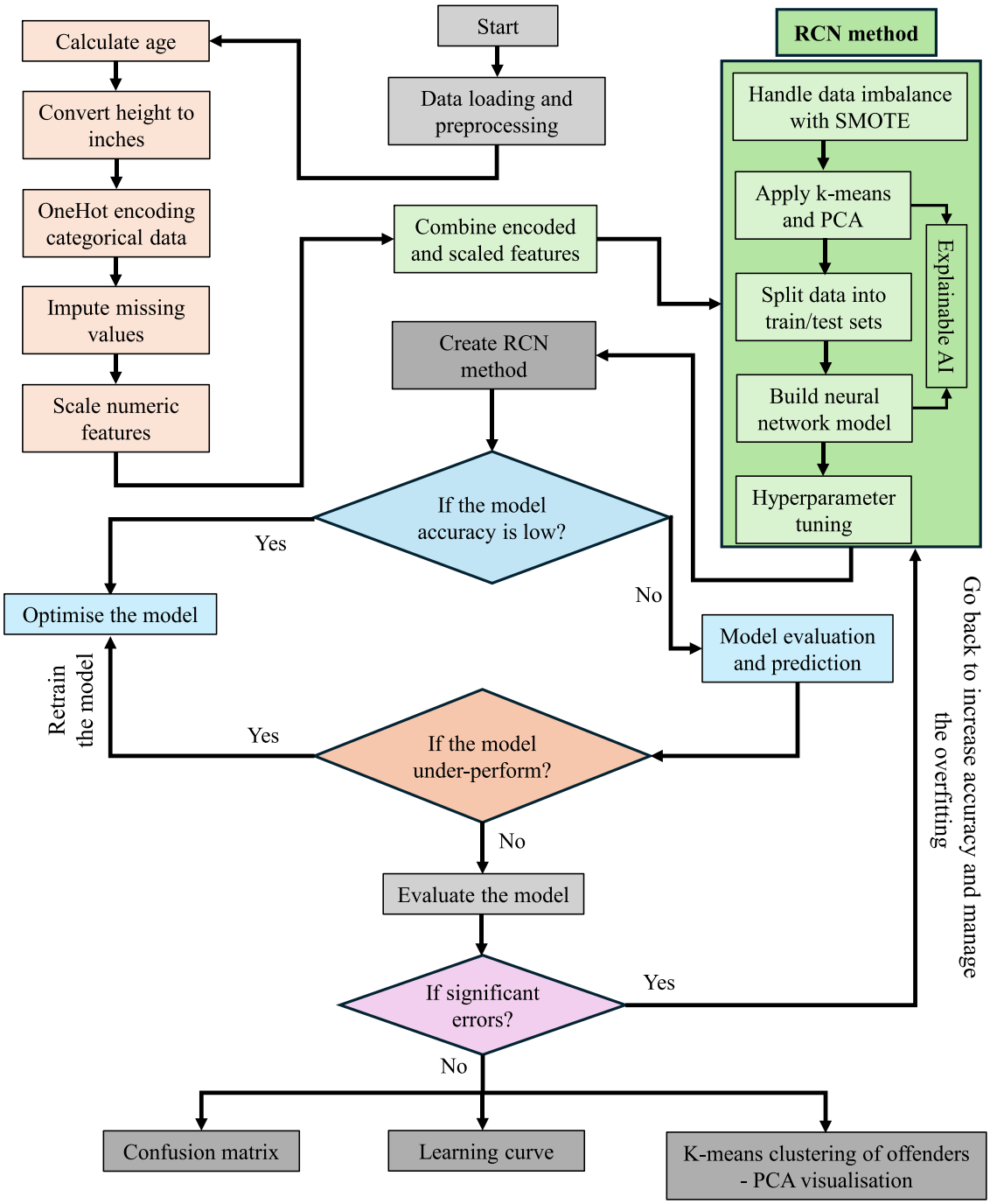


Fig. 4. Flow chart of our proposed method.

4.4. Methodological steps for implementing the RCN system

The flowchart in Fig. 4 illustrates the comprehensive process of the RCN method, designed to predict and analyze the likelihood of recidivism, or re-offending, among individuals. The process begins with data loading and preprocessing. It involves preparing the raw dataset by handling missing values, encoding categorical variables, scaling numeric features, and computing important attributes like the age of individuals from their date of birth. Preprocessing also includes converting height from feet and inches into inches for consistency and applying techniques like one-hot encoding for categorical data. These initial steps ensure that the data is in the right format and scale to be processed by the ML model.

Once the data is preprocessed, the next step is handling data imbalance using SMOTE. Since the dataset may have fewer instances of recidivism (offenders) compared to non-offenders, SMOTE helps balance the data by generating synthetic samples for the minority class, making the model more effective in identifying recidivism cases. After this, the encoded categorical and scaled numeric features are combined into a single dataset, preparing the data for input into the model. The process then leads to the creation of the RCN method, where additional techniques, such as k-means clustering and PCA, are used to reduce the dimensionality of the data, making it easier to visualize and understand key patterns within the offender population. The next stage involves building the neural network model. The dataset is split into training and testing sets to evaluate the model's performance. A

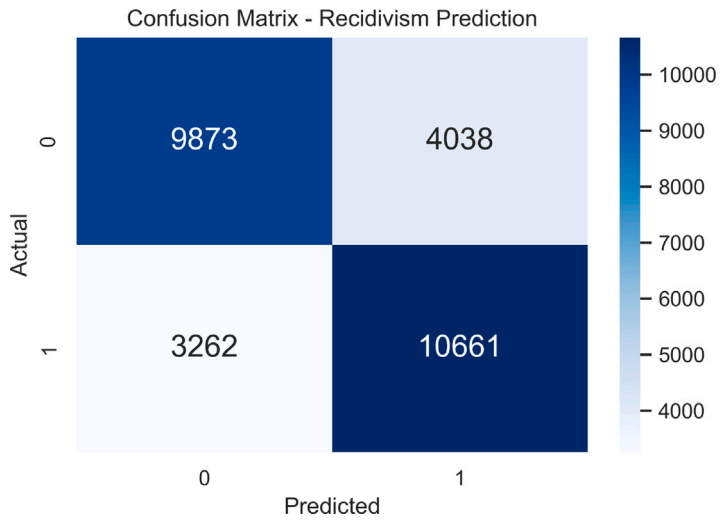


Fig. 5. Confusion matrix for recidivism prediction.

DL model, specifically a neural network, is built to predict recidivism. This neural network undergoes hyper-parameter tuning to optimize its performance. During this step, key parameters such as learning rate, number of layers, and units in the neural network are adjusted to achieve the best possible results. Moreover, the model includes an XAI component, ensuring the predictions are interpretable and transparent. This is critical in legal contexts where the decision-making process must be understood and justified.

After training the model, it is evaluated for accuracy and prediction performance. If the model's accuracy is low, it enters a loop where it is optimized through further tuning and retraining until satisfactory results are obtained. In cases where the model underperforms or significant errors are detected, adjustments are made to improve the model's performance. The iterative optimization process ensures that the final model is robust and accurate. Once the model is optimized and performs well, the final step involves visualizing the model. Key outputs such as the confusion matrix and learning curve provide insight into the model's performance. The confusion matrix helps assess how well the model distinguishes between true positives, true negatives, false positives, and false negatives, giving a clear picture of its predictive accuracy. The learning curve shows how the model's performance improves over time, helping detect overfitting or underfitting issues. Additionally, the k-means clustering visualization using PCA is generated to display the offender clusters. This visualization reveals patterns in the offender population, helping to identify groups with similar features and their likelihood of recidivism. In conclusion, the flowchart represents a systematic approach to building and refining an ML model for predicting recidivism. The process involves careful data preprocessing, handling imbalances, model development, and optimization, ensuring the predictions are explainable and interpretable. Through this structured approach, the RCN method effectively identifies patterns and provides insights into offender behavior, which can be crucial for legal and correctional systems.

5. Evaluating RCN performance and insights from offender profiling

This section presents the results of our analysis using the RCN method. The results are divided into model performance metrics and offender clustering using dimensionality reduction techniques. Figures visualizing the relationships between offender characteristics and recidivism tendencies and the prediction model's performance support the findings.

5.1. Model performance

The first part of our results focuses on the performance of the recidivism prediction model, evaluated using metrics such as the confusion matrix, learning curves for accuracy and loss, precision–recall curve, and ROC curve. These metrics provide a comprehensive view of the model's effectiveness in predicting recidivism.

- Confusion matrix: The confusion matrix in Fig. 5 provides a detailed overview of the model's performance in predicting recidivism. The model successfully classified 10,661 individuals as recidivists (True Positives) and 9,873 individuals as non-recidivists (True Negatives), demonstrating its ability to identify a large proportion of correct cases. However, the matrix also reveals that 4,038 non-recidivists were misclassified as recidivists (False Positives), and 3,262 recidivists were misclassified as non-recidivists (False Negatives). These misclassifications highlight a trade-off between sensitivity and specificity, where the model captures many true positives but also produces a notable number of false positives and false negatives. Despite these limitations, the overall performance is encouraging, with 20,534 out of 25,834 instances classified correctly. This suggests that the model has the potential for predictive accuracy in recidivism forecasting, though further refinement is necessary to reduce misclassification rates and improve the balance between precision and recall.
- Learning curves for accuracy: The accuracy curve shows the model's classification performance. Both training and validation accuracy increase over epochs, indicating improvement in the model's predictive capabilities. The training accuracy reaches approximately 0.76 by epoch 50 in Fig. 6(a), while validation accuracy stabilizes around 0.72–0.74. This consistent gap suggests that while the model is performing well on the training data, it also maintains a similar level of performance on the validation data, showing good generalization with minimal overfitting.
- Learning curve - loss: The loss curve reflects how well the model fits the training data compared to its generalization on the validation data. During the initial epochs (0–10) in Fig. 6(b), both training and validation loss decrease rapidly, indicating effective learning by the model. As training progresses, the training loss continues to decrease steadily, reaching about 0.48 by epoch 50, which suggests a better fit for the training data. The validation loss decreases initially but stabilizes around 0.52–0.55, with some minor fluctuations. The small gap between training and validation loss toward the end suggests the model fits the training data slightly better but is not significantly overfitting, as

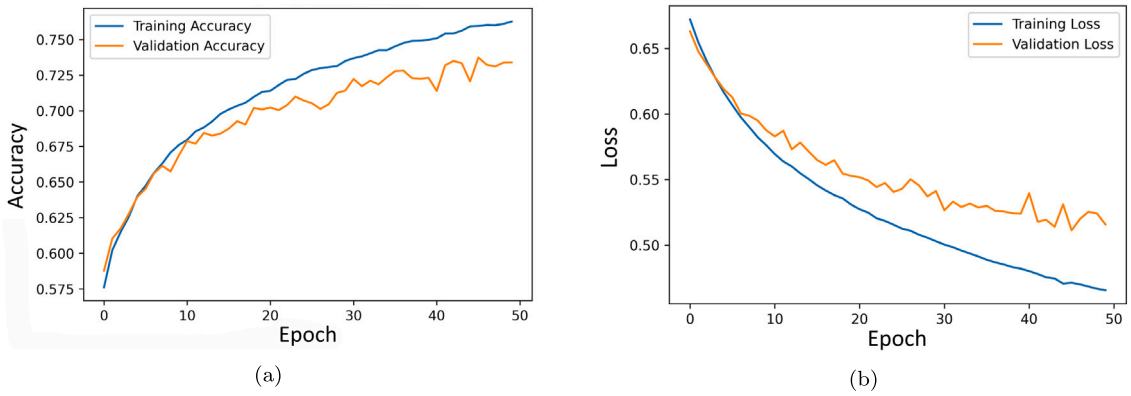


Fig. 6. Learning curves for (a) accuracy and (b) loss.

both losses remain relatively close. Overall, these curves indicate a well-trained model with a balanced trade-off between fitting the training data and generalizing to validation data. The small performance gap suggests minor room for further tuning, but the model demonstrates stable learning with minimal signs of overfitting.

5.2. Offender clustering and dimensionality reduction

To gain further insights into the patterns within the offender population, we applied clustering and dimensionality reduction techniques, including PCA and t-SNE, to visualize the latent groupings among offenders. As presented in Fig. 7, the pairwise scatter plot comprehensively examines the interrelations between key offender characteristics—age, height, weight, and time since the last conviction—across the four identified clusters. These clusters, distinguished by different colors (Cluster 0 in purple, Cluster 1 in green, Cluster 2 in yellow, and Cluster 3 in blue), exhibit varying patterns of feature distribution and relationships.

- Age distribution: The kernel density estimate (KDE) plots on the diagonal show significant differences in age distributions across clusters. Cluster 2 (yellow) is characterized by a notably younger population, with most individuals in this cluster being below 50 years of age. In contrast, Cluster 0 (purple) and Cluster 1 (green) display wider age distributions. Cluster 0, in particular, includes a substantial number of older individuals, some exceeding 80 years of age. This cluster seems to represent a more diverse age group, suggesting that it includes offenders with varying lengths of criminal history.
- Height and weight: Height and weight reveal interesting clustering dynamics. Cluster 0 (purple) and Cluster 3 (blue) tend to concentrate around mid-range values for both height (approximately 60–75 inches) and weight (around 150–250 pounds), as evidenced by the clustering of points in these regions. Cluster 2 (yellow), which consists of younger individuals, appears more dispersed across the weight spectrum, with some offenders weighing between 100 and 400 pounds. In terms of height, this cluster shows a broader spread, though with a heavier representation at the lower end of the height scale (50–60 inches). The KDE plots confirm that Cluster 0 (purple) has a narrower height distribution, indicating that this group has less variability in stature compared to other clusters, particularly Cluster 2 (yellow), which exhibits more diversity.
- Time since last conviction: Time since last conviction appears strongly correlated with age across all clusters. Cluster 2 (yellow), representing younger individuals, shows a concentration of offenders with shorter times since their last conviction, often less than 30 years. This suggests that this group has more recent criminal activities. On the other hand, Cluster 0 (purple) and Cluster 1 (green), which include older individuals, have longer periods since their last conviction, often between 40 and 70 years. This could imply that these clusters represent offenders who have either aged out of crime or have had a long history of recidivism followed by more extended periods without re-offending. The KDE plot for time since the last conviction further highlights that Cluster 0 (purple) encompasses a wider range of recidivism histories, with a higher proportion of offenders who have not re-offended for an extended period.
- Interrelationships between features: The scatter plots in Fig. 7 between pairs of features offer additional insights. For instance, the relationship between age and weight shows some degree of linearity across clusters, particularly within Cluster 0 (purple), where older offenders tend to have higher weights. This may reflect physiological changes or lifestyle factors that differentiate older offenders. Cluster 2 (yellow) exhibits a more scattered pattern with no clear linear relationship, indicating that younger offenders in this cluster vary widely in weight regardless of age. Height does not show strong correlations with other features, suggesting that it may not be a significant predictor in distinguishing offender clusters.
- Cluster characteristics: Each cluster represents a distinct offender profile. Cluster 2 (yellow), consisting mainly of younger and more recently convicted offenders, could represent individuals at a higher risk of re-offending in the near future. Cluster 0 (purple), which includes a broader age range and longer times since the last conviction, may consist of offenders who have aged out of crime or have successfully reintegrated into society after long periods of criminal behavior. Cluster 1 (green) and Cluster 3 (blue) seem to represent intermediate profiles with varied characteristics but less extreme distributions in terms of age and time (0 to 3). Specifically, the clustering pattern in the PCA space indicates that distinct offender groups emerge based on the selected features. Cluster 0 (dark purple) and Cluster 2 (yellow) are well-separated, suggesting these clusters represent groups of offenders with significantly different feature profiles. On the other hand, Clusters 1 (green) and 3 (blue) overlap somewhat, indicating these clusters may share certain features or exhibit more nuanced differences. Quantitatively, these results suggest that the PCA transformation captures the underlying structure of the data, with the first two components explaining a considerable portion of the variance. This visualization aids in interpreting the offender groupings, but further analysis of the underlying feature contributions to each cluster is necessary to understand the behavioral or demographic characteristics driving these separations.

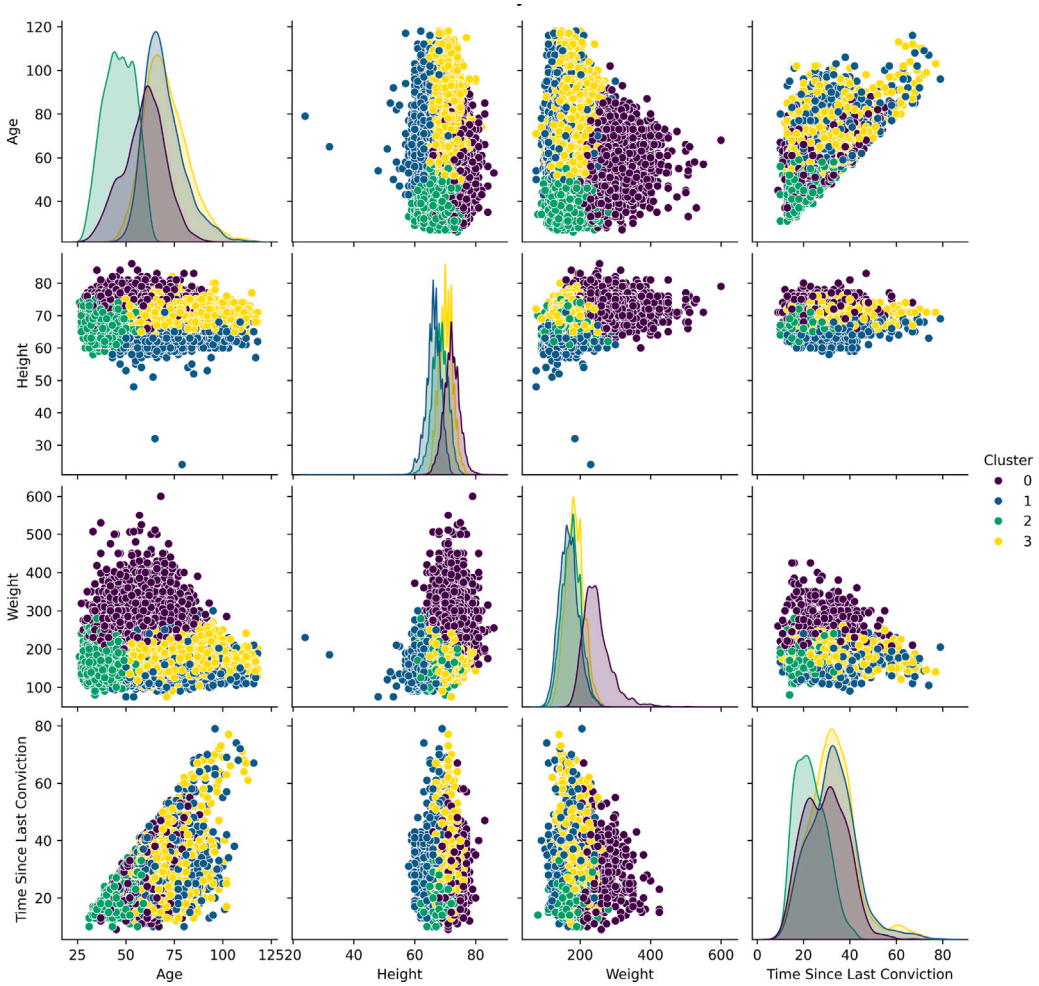


Fig. 7. Pairwise scatter plot of key offender features across clusters.

5.3. PCA clustering visualization with two principal components

The PCA visualization in Fig. 8 presents a two-dimensional reduction of the offender clusters, demonstrating distinct group separations. The k-means clustering has segmented the data into four clusters, each clearly distinguished within the feature space. Cluster 0 (purple) is concentrated in the lower-left region, spanning approximately from −7.5 to −2.5 on the first principal component and −3 to 0 on the second component. Cluster 1 (yellow), located in the upper right, extends from 0 to 5 on the first component and from 0 to 4 on the second. Cluster 2 (green) and Cluster 3 (blue) overlap in the central region, with the former spanning from −2.5 to 2.5 on the first component and −2 to 3 on the second, while the latter is more central, ranging from −5 to 2.5 on the first component and from −2 to 2 on the second. The first principal component explains a significant portion of variance in the dataset, extending from −7.5 to 5, while the second principal component captures variance along the vertical axis from −3 to 4. This distinct clustering indicates that offenders within each group share similar attributes, such as age, height, weight, and recidivism characteristics, suggesting varying recidivism risks. These clearly separated clusters reflect underlying patterns in offender behavior and demographics, emphasizing the effectiveness of PCA in preserving essential differences while reducing the dataset's dimensionality.

5.4. t-SNE visualizations

In Fig. 9, we see six t-SNE visualizations that illustrate clusters of offenders using different settings for perplexity and iteration parameters, allowing us to observe how these adjustments impact the distribution and clarity of clusters in offender data. Each subfigure uses a different combination of perplexity and iteration values, from lower (5) to higher (50) perplexity levels and with iteration counts of 500 or 1000.

Fig. 9(a) (Perplexity=5, Iterations=500) provides a focused view on local groupings, capturing clusters that are more concentrated and less influenced by global data relationships. The model primarily reflects neighborhood relationships at this low perplexity, resulting in several small, well-defined clusters with minimal overlap. Fig. 9(b) (Perplexity=5, Iterations=1000) shows a similar cluster structure but slightly more refined boundaries due to the increased number of iterations. The low perplexity of these two visualizations highlights specific, localized groupings of offenders, suggesting that offenders have highly similar characteristics within these clusters. On the one side, Fig. 9(c) (Perplexity=30, Iterations=500) and Fig. 9(d) (Perplexity=30, Iterations=1000) represent an intermediate level of perplexity. This setting balances the focus between local and global relationships, capturing distinct and broad clusters. In Fig. 9(c), the clusters remain clearly separated with some spread, while Fig. 9(d) benefits from additional iterations to delineate cluster edges further. These intermediate perplexity settings allow us to observe patterns that consider more distant relationships, which may indicate similarities across offender groups beyond immediate neighborhood connections. This balance helps identify groups of offenders who share broader commonalities, not just localized ones. Finally, Fig. 9(e) (Perplexity=50, Iterations=500) and Fig. 9(f) (Perplexity=50, Iterations=1000) use a high perplexity setting to emphasize global structures within the data. Here, the clusters are larger and overlap, reflecting a broader perspective on the relationships within the

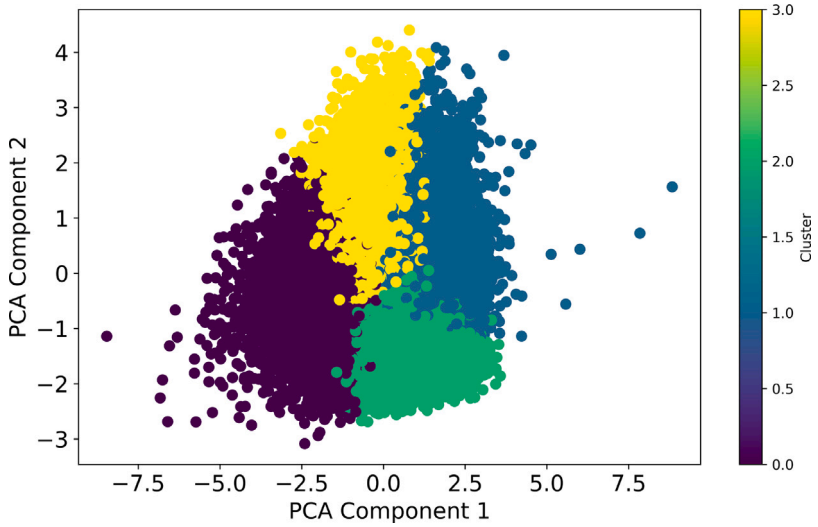


Fig. 8. PCA clustering of offenders based on recidivism-related features.

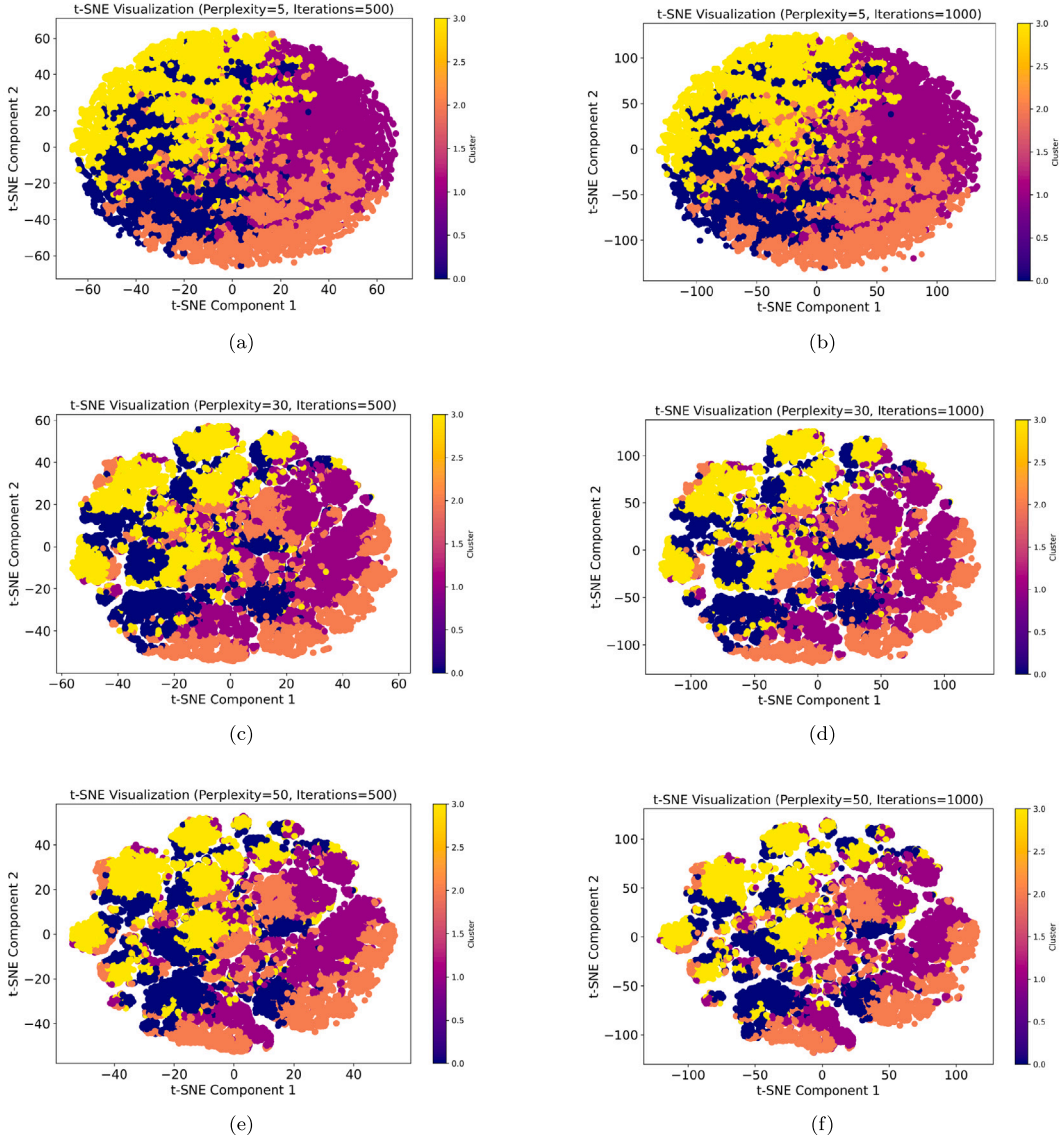


Fig. 9. Comparison of t-SNE clustering results under various perplexity and iteration settings: (a) Perplexity=5, Iterations=500; (b) Perplexity=5, Iterations=1000; (c) Perplexity=30, Iterations=500; (d) Perplexity=30, Iterations=1000; (e) Perplexity=50, Iterations=500; (f) Perplexity=50, Iterations=1000.

data, where neighborhood distinctions are less prominent. In Fig. 9(e), we observe more general cluster groupings, while in Fig. 9(f), increasing the iteration count to 1000 helps to sharpen the separation between clusters, although some overlap remains. These settings illustrate how higher perplexity aggregates individuals based on distant similarities, capturing more of the overall data structure rather than focusing on immediate relationships.

Overall, the figures demonstrate how t-SNE parameters can shape the visualization of clusters in offender data. Lower perplexity settings (Figs. 9(a) and 9(b)) highlight specific local patterns, intermediate perplexity (Figs. 9(c) and 9(d)) reveals balanced clusters that capture both local and global aspects, while higher perplexity settings (Figs. 9(e) and 9(f)) provide a broader view of offender groupings. Adjusting these parameters allows for a tailored view of the data, helping to reveal insights into how offenders might be grouped based on both close and distant relationships in terms of recidivism risk factors.

When the number of iterations is increased to 1000 while keeping perplexity at 5, the clusters become more distinct, yet the global structure is still more prominent than the fine-grained local relationships. This visualization illustrates how longer iterations can enhance the clarity of clustering even when using lower perplexity values. The clustering patterns balance local and global structures at a perplexity of 30 with 500 iterations. The clusters are distinct yet moderately dispersed, showing that this configuration captures both fine details and broader relationships. The clusters maintain cohesion but with slightly more spread than at higher perplexity values. Increasing the number of iterations to 1000 at a perplexity of 30 refines the clusters further, resulting in more compact groupings. The clusters are well-separated, with sharp transitions, indicating that the increased iterations help achieve a more precise data structure resolution. Finally, using perplexity of 50 and 500 iterations, the clusters remain distinct but slightly less compact compared to higher iteration counts. The general structure of the clusters is still well-captured, though the borders between clusters are less defined, suggesting that higher iterations improve the overall refinement of the clustering. These t-SNE visualizations highlight the impact of tuning perplexity and iterations on the clarity and precision of clustering. Lower perplexity values tend to reveal broader global structures, while higher perplexity focuses more on local relationships. Increasing iterations generally results in more refined clusters, enhancing the visual distinction between groups within the dataset. This analysis provides a nuanced understanding of the underlying patterns within the data, offering valuable insights for further exploration and interpretation.

5.5. Precision-recall trade-offs in recidivism prediction

Balancing precision and recall in judicial contexts is crucial due to the ethical implications of false positives and negatives. The trade-off between these metrics significantly affects decision-making in criminal justice, influencing sentencing, parole approvals, and rehabilitation programs.

False Positives (FP): A high false positive rate implies that individuals unlikely to reoffend are incorrectly classified as high-risk. This can lead to excessive sentencing, parole denials, and unnecessary monitoring, raising concerns about fairness and unjust restrictions [19]. Over-reliance on high recall models may contribute to systemic biases, disproportionately affecting marginalized communities.

False Negatives (FN): A high false negative rate means that actual recidivists are misclassified as low-risk, potentially endangering public safety. While a high-precision model reduces false positives, it may increase false negatives, allowing individuals with a high probability of reoffending to be released into society [20]. This highlights the necessity of balancing these trade-offs to ensure both fairness and security.

Ethical Considerations: Achieving an optimal precision–recall balance requires carefully calibrating risk assessment models. Techniques such as cost-sensitive learning, fairness-aware algorithms, and threshold tuning can help mitigate biases while maintaining predictive accuracy [23]. Moreover, policymakers and legal experts must be involved in setting ethical thresholds, ensuring transparency and accountability in algorithmic decision-making.

Researchers can design models that minimize harm while maintaining judicial fairness by considering precision–recall trade-offs. Future work should explore dynamic threshold adjustments and adversarial debiasing techniques to refine predictive fairness in recidivism assessment.

6. Interpreting findings and ethical considerations for AI in criminal justice

The results demonstrate the utility of the RCN method in identifying patterns within offender populations and predicting recidivism with a moderate degree of accuracy. The feature distribution analysis suggests that age and time since the last conviction may be relevant to recidivism outcomes but are not strong standalone predictors. Instead, combining these features with clustering techniques can reveal more complex relationships that might otherwise remain hidden in higher-dimensional spaces.

The confusion matrix in Fig. 5 demonstrates the model's ability to distinguish between recidivists and non-recidivists. While the model correctly classifies a substantial number of instances, there remains a notable proportion of false positives (4,038) and false negatives (3,262), as reflected in the confusion matrix. These misclassifications underscore the inherent complexity of predicting recidivism, where the trade-off between sensitivity and specificity presents ongoing challenges. Addressing these false predictions remains critical for enhancing the model's reliability and ensuring its practical application in recidivism forecasting. Despite these misclassifications, the model effectively captures a significant number of true positives and true negatives, indicating moderate success in predicting recidivism. On the one side, the learning curves for accuracy and loss provide insight into the training process in Fig. 6. The training accuracy steadily improves, with validation accuracy following a similar trend, though with some fluctuation due to overfitting. The training loss also decreases progressively, while validation loss shows more variance, reflecting the model's difficulty in generalizing across unseen data. This suggests that while the model is learning the underlying patterns in the training data, it has challenges when applied to new data.

The clustering analysis (Figs. 8 and 9 using PCA and t-SNE visualizations sheds light on the complex structure of the dataset. The PCA visualization shows that the clusters are fairly distinct when reduced to two dimensions, though there is some overlap between clusters. The pairwise scatter plot of key features across clusters also supports this observation, showing relationships between features such as age, height, and weight, with clear separation in some areas and overlap in others. This demonstrates the utility of clustering methods in uncovering latent structures within the data, though the complexity of the offender profiles creates challenges in achieving perfectly distinct clusters. The t-SNE visualizations across different perplexities (5, 30, 50) and iterations (500, 1000) provide further insights into the clustering structure. Lower perplexities produce more tightly grouped clusters, while higher perplexities lead to more spread-out groupings. As the number of iterations increases, the clusters become more refined, though some overlap remains between different clusters. These visualizations suggest that t-SNE effectively captures the non-linear relationships between features, revealing the nuanced distribution of data points within the two-dimensional space. However, the presence of overlapping clusters highlights the inherent complexity of the dataset. It suggests that further feature engineering or advanced modeling techniques may be required to improve cluster separation. Overall, the results indicate that while the model and clustering methods provide valuable insights into the recidivism dataset, significant challenges

remain in accurately predicting recidivism and fully disentangling the complex relationships between offender attributes. The findings point to the need for continued refinement of both predictive models and clustering algorithms, particularly in terms of improving generalization and cluster separation.

The RCN method presents an effective approach for offender profiling and recidivism prediction, combining DL, clustering, and XAI techniques. This approach can potentially provide more interpretable and actionable insights, aiding decision-making processes in the criminal justice system. Future work should explore the potential for expanding the model to include a broader range of variables and more sophisticated analysis methods to enhance predictive power and fairness.

7. Conclusions

This study introduced the RCN, an effective method that integrates deep DL, clustering techniques, and XAI to enhance the prediction of recidivism and offender profiling. The RCN method combines ML models optimized with the Keras tuner. Clustering approaches include k-means, dimensionality reduction techniques, and PCA and t-SNE. This approach allowed for accurate recidivism predictions and the identification of latent offender clusters, offering deeper insights into the behavioral patterns contributing to recidivism. The model achieved a reasonable accuracy of nearly 75%, correctly identifying 10,661 recidivists, though 4,038 false positives and 3,262 false negatives were also recorded. This result reflects the challenges inherent in balancing sensitivity and specificity, emphasizing the need for further model refinement to reduce misclassifications, particularly false positives and false negatives. The t-SNE visualizations proved valuable in revealing distinct clusters of offenders, demonstrating that certain subgroups with shared characteristics may have different recidivism tendencies. Furthermore, by incorporating SHAP values, the RCN method offered a high level of model interpretability, making the predictions more transparent and understandable for stakeholders within the criminal justice system. This transparency is crucial in ensuring that AI-driven decisions are ethical and justifiable. While the results are promising, future research should improve the model by expanding the feature set to include more socio-economic, psychological, and behavioral factors that could enhance its predictive power. Exploring advanced ML techniques, such as ensemble models, could further refine performance. Overall, the RCN method represents a significant advancement in the use of AI in criminal justice, providing a robust and interpretable tool to support more informed judicial decision-making and to help shape targeted interventions for reducing recidivism.

CRediT authorship contribution statement

Muhammed Cavus: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Muhammed Nurullah Benli:** Writing – review & editing, Methodology. **Usame Altuntas:** Writing – review & editing, Methodology. **Mahmut Sari:** Writing – review & editing, Resources. **Huseyin Ayan:** Writing – review & editing, Methodology. **Yusuf Furkan Ugurluoglu:** Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was funded by the Turkish Ministry of National Education.

Data availability

The code and data are publicly available on GitHub at https://github.com/cavusmuhammed68/Recidivism-Clustering-Network-RCN-

References

[1] M. Schmucker, F. Lösel, The effects of sexual offender treatment on recidivism: An international meta-analysis of sound quality evaluations, J. Exp. Criminol. 11 (2015) 597–630.
[2] R.K. Bharati, Ethical implications of AI in criminal justice: Balancing efficiency and due process, Res. Rev. Int. J. Multidiscip. 9 (7) (2024) 93–105.
[3] M. Miron, S. Tolan, E. Gómez, C. Castillo, Evaluating causes of algorithmic bias in juvenile criminal recidivism, Artif. Intell. Law 29 (1) (2021) 111–112.
[4] C. Engel, L. Linhardt, M. Schubert, Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism, Artif. Intell. Law (2024) 1–23.
[5] M. Portela, C. Castillo, S. Tolan, M. Karimi-Haghighi, A.A. Pueyo, A comparative user study of human predictions in algorithm-supported recidivism risk assessment, Artif. Intell. Law (2024) http://dx.doi.org/10.1007/s10506-024-09393-y.
[6] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European Court of Human Rights, Artif. Intell. Law 28 (2) (2020) 237–266.
[7] G.V. Travaini, F. Pacchioni, S. Bellumore, M. Bosia, F. De Micco, Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction, Int. J. Environ. Res. Public Heal. 19 (17) (2022) 10594, http://dx.doi.org/10.3390/ijerph191710594.
[8] N. Tollenaar, B. Wartna, P. Van Der Heijden, S. Bogaerts, StatRec-performance, validation and preservability of a static risk prediction instrument, Bull. Sociol. Methodol./Bull. de Méthodol. Sociol. 129 (1) (2016) 25–44.
[9] M. Van Hall, T. Baker, A.J. Dirkzwager, P. Nieuwbeerta, Perceptions of probation officer procedural justice and recidivism: A longitudinal study in the Netherlands, Crim. Justice Behav. (2024) 00938548241244502.
[10] J. Meijers, J.M. Harte, G. Meynen, P. Cuijpers, Differences in executive functioning between violent and non-violent offenders, Psychol. Med. 47 (10) (2017) 1784–1793.
[11] S. Guo, Y. Wang, Investigating predictors of juvenile traditional and/or cyber offense using machine learning by constructing a decision support system, Comput. Hum. Behav. 152 (2024) 108079.
[12] G. Gao, K. Xiao, H. Li, S. Song, An intelligent assessment method of criminal psychological attribution based on unbalance data, Comput. Hum. Behav. 158 (2024) 108286.
[13] N. Tollenaar, P.G.M. van der Heijden, Which method predicts recidivism best?: A comparison of statistical, machine learning and data mining predictive models, J. R. Stat. Soc. Ser. A: Stat. Soc. 176 (2) (2013) 565–584, http://dx.doi.org/10.1111/j.1467-985X.2012.01056.x.
[14] G. Duwe, K. Kim, Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism, Crim. Justice Policy Rev. 28 (6) (2017) 570–600, http://dx.doi.org/10.1177/0887403415604899.
[15] J. Caulkins, J. Cohen, W. Gorr, J. Wei, Predicting criminal recidivism: A comparison of neural network models with statistical methods, J. Crim. Justice 24 (3) (1996) 227–240, http://dx.doi.org/10.1016/0047-2352(96)00012-8.
[16] M.H. Ting, C.M. Chu, G. Zeng, D. Li, G.S. Chng, Predicting recidivism among youth offenders: Augmenting professional judgement with machine learning algorithms, J. Soc. Work. 18 (6) (2018) 631–649, http://dx.doi.org/10.1177/1468017317743137.
[17] M.L. Jamil, S. Pais, N. Pombo, J. Cordeiro, P. Neves, Challenges of profiling offenders for recidivism risk, in: 2023 IEEE International Conference on Data Mining Workshops, ICDMW, IEEE, 2023, pp. 235–244.
[18] V. Mandalapu, L. Elluri, P. Vyas, N. Roy, Crime prediction using machine learning and deep learning: A systematic review and future directions, IEEE Access 11 (2023) 60153–60170.
[19] R. Berk, Criminal Justice Forecasts of Risk: A Machine Learning Approach, Springer Science & Business Media, 2012.
[20] G. Duwe, K. Kim, Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance, Corrections 1 (3) (2016) 155–176.
[21] T. Ozkan, Predicting Recidivism Through Machine Learning (Ph.D. thesis), University of Texas at Dallas, 2017, Digital access provided by the Eugene McDermott Library URL http://hdl.handle.net/10735.1/5405.
[22] R. de la Cruz, O. Padilla, M.A. Valle, G.A. Ruz, Modeling recidivism through Bayesian regression models and deep neural networks, Mathematics 9 (6) (2021) 639.
[23] K.T. Rodolfa, E. Salomon, L. Haynes, I.H. Mendieta, J. Larson, R. Ghani, Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 142–153.
[24] J. Ryberg, Criminal justice and artificial intelligence: How should we assess the performance of sentencing algorithms? Philos. Technol. 37 (2024) 9.

[25] R. Dipshan, V. Hudgins, F. Ready, The United States of risk assessment: The machines influencing criminal justice decisions, 2020, https://www.law.com/legaltechnews/2020/07/13/the-united-states-of-risk-assessment-the-machines-influencing-criminal-/justice-decisions/?slreturn=20241016104627.
[26] G. Zara, D.P. Farrington, Criminal Recidivism: explanation, Prediction and Prevention, Routledge, 2016.
[27] S. Anwar, J. Engberg, I.M. Opper, L. Dion, What Happens When Judges Follow the Recommendations of Pretrial Detention Risk Assessment Instruments More Often?, Tech. Rep., RAND, 2024.
[28] J. Zhang, Research on the criminal recidivism prediction based on machine learning algorithm, in: Proceedings of the 2022 2nd International Conference on Business Administration and Data Science (BADS 2022), Atlantis Press International BV, Dordrecht, 2023, pp. 1297–1306, http://dx.doi.org/10.2991/978-94-6463-102-9_134.
[29] G. van Dijck, Predicting recidivism risk meets AI act, Eur. J. Crim. Policy Res. 28 (3) (2022) 407–423, http://dx.doi.org/10.1007/s10610-022-09516-8.
[30] D. Mu, S. Zhang, T. Zhu, Y. Zhou, W. Zhang, Prediction of recidivism and detection of risk factors under different time windows using machine learning techniques, Soc. Sci. Comput. Rev. (2024) http://dx.doi.org/10.1177/08944393241226607.
[31] V. Li, S. Sridharan, S. Sethuraman, G. Avdis, Predicting recidivism with machine learning: An analysis of risk factors and proposal of preventions, J. Stud. Res. 12 (4) (2023) http://dx.doi.org/10.47611/jsrhs.v12i4.5779.
[32] M. Karimi-Haghighi, C. Castillo, Enhancing a recidivism prediction tool with machine learning, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ACM, New York, NY, USA, 2021, pp. 210–214, http://dx.doi.org/10.1145/3462757.3466150.
[33] R.L. Heilbronner, Lockett v. Ohio (1978), in: Encyclopedia of Clinical Neuropsychology, Springer New York, New York, NY, 2011, pp. 1479–1480, http://dx.doi.org/10.1007/978-0-387-79948-3_1003.
[34] V. Woodson, North carolina, 428 US 280, U. S. Supreme Court. (1976).
[35] W.W. Berry, Individualized sentencing, Wash. Lee Law Rev. 76 (1) (2019) 19.
[36] N. Sabouret, Understanding Artificial Intelligence, CRC Press, 2020.
[37] Wisconsin Supreme Court, State v. Loomis, North West. Report. Second. Ser. 881 (2016) 749–761, Court Decision. URL https://www.leagle.com/decision/inwico20160713i48.
[38] H.-W. Liu, C.-F. Lin, Y.-J. Chen, Beyond State v Loomis: artificial intelligence, government algorithmization and accountability, Int. J. Law Inf. Technol. 27 (2) (2019) 122–141.
[39] A. Sachoulidou, Going beyond the "common suspects": To be presumed innocent in the era of algorithms, big data, and artificial intelligence, Artif. Intell. Law (2023) http://dx.doi.org/10.1007/s10506-023-09347-w.
[40] R.V. Zicari, J. Amann, F. Bruneault, M. Coffee, B. Düdder, E. Hickman, A. Gallucci, T.K. Gilbert, T. Hagendorff, I. van Halem, et al., How to assess trustworthy AI in practice, 2022, arXiv preprint arXiv:2206.09887.
[41] E.U. Artificial Intelligence Act, The EU artificial intelligence act, 2024.
[42] Gardner v. Florida, Gardner v. Florida 430 U.S. 349, 1977, (U.S. Supreme Court).
[43] E.L. LOOMIS, Loomis v. Wisconsin, 2016, pp. 749–765, URL https://www.leagle.com/decision/inwico20160713i48.
[44] European Commission, Ethics guidelines for trustworthy AI, 2019, Available at https://ec.europa.eu/digital-strategy/our-policies/european-approach-artificial-intelligence_en.
[45] OECD, OECD principles on artificial intelligence, 2019, Available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
[46] White House Office of Science and Technology Policy, Blueprint for an AI bill of rights, 2022, Available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
[47] D.K. Citron, F. Pasquale, The scored society: Due process for automated predictions, Wash. Law Rev. 89 (1) (2014) 1–33.
[48] M. Hamilton, P. Ugwudike, A 'black box' AI system has been influencing criminal justice decisions for over two decades–it's time to open it up, 2023.
[49] M. Oswald, J. Grace, S. Urwin, G.C. Barnes, Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'experimental' proportionality, Inf. Commun. Technol. Law 27 (2) (2018) 223–228, http://dx.doi.org/10.1080/13600834.2018.1458455.
[50] F. Lütz, The AI act, gender equality, and non-discrimination: What role for the AI office? ERA Forum 25 (2024) 79–95, http://dx.doi.org/10.1007/s12027-023-00709-2.
[51] A.R. Lewis, M.R. Rief, M.D. Applegarth, Best Practices for Improving the Use of Criminal Justice Risk Assessments: Insights from NIJ's Recidivism Forecasting Challenge Winners Symposium, National Institute of Justice, 2024, https://nij.ojp.gov/topics/articles/best-practices-improving-use-criminal-justice-risk-assessments.
[52] J.D. Jackson, S.J. Summers, The presumption of innocence, in: The Internationalisation of Criminal Evidence: Beyond the Common Law and Civil Law Traditions (Law in Context), Cambridge University Press, 2012, p. 200.
[53] W. Blackstone, Commentaries on the Laws of England (Book IV), Clarendon Press, pp. 1765–1769.
[54] California Department of Justice, Megan's law, retrieved from https://www.meganslaw.ca.gov (n.d.).