International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05680
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Behaviour-Aware Explainable Artificial Intelligence Framework for Criminal Offender Profiling

1. Dr.K.Priya
Assistant Professor,
Department of Artificial Intelligence and Machine Learning,
SoISDS, KPRCAS, Coimbatore

2. Dr.K.Dheenathayalan
Assistant Professor, Department of Computer Science
NIFT-TEA College of Knitwear Fashion, Tirupur

Abstract: In recent years the use of artificial intelligence for criminal offender profiling has grown which at the same time has seen the2 issue of transparency in most predictive models which has not gone away still we have issues with belief in these models, with who is responsible when something goes wrong, and also the ethical play out of these in the criminal justice field. We don\'t see it as sufficient that a model is accurate in criminal justice settings; what we also require is that the model's decisions are also made clear to human stakeholders. That is what this paper sets out to do which is present an Explainable AI based framework in which we look at the behavioural patterns of the offender at risk instead of the dense stats. Also, we put forth a set of behaviour-based features which we get from past offense reports. Recidivism risk is estimated using ensemble-based machine learning models, and explainability techniques are then used to interpret the predictions. These explanations offer both localised explanations for specific offender predictions and global perceptions of important behavioural factors. According to the experimental results, the suggested framework significantly improves model interpretability while achieving competitive predictive performance.

Keywords: Explainable Artificial Intelligence, Criminal Offender Profiling, Recidivism Prediction, Behavioural Analytics, SHAP and LIME.

1. Introduction

More and more artificial intelligence (AI) systems have been introduced into the criminal justice system in recent years to assist with decision making processes, including risk assessment, sentencing, parole and offender classification. Predictive models based on machine learning lend themselves to sift through vast swaths of criminal history data and help identify patterns that may not be as apparent to a human decision maker. Hence automated risk assessment tools have been identified as having the potential to increase consistency and efficiency in criminal justice proceedings.
One corresponding trade-off of increased reliance on AI-driven systems, though, is that the transparency, accountability and ethical deployment are significant concerns. There are a lot of hair-raising assumptions that affect the accuracy of these black-box models, including issues related to sampling, machine learning, and spurious correlations. This is particularly problematic in high-stakes settings - such as criminal justice - where automated decisions can have a large effect on personal freedom, legal consequences, and public trust. As a result, much attention has recently been focused on explainable AI (XAI) approaches that permit human stakeholders to interpret, inspect and justify predictions of models.

Recidivism is instrumental in offender profiling and criminal risk assessment. Recidivism is the subsequent reoffending of an individual, as indicated by criminal behaviour that results in rearrests, reconviction or reimprisonment. The ability to predict risk of reoffending can help policy makers and practitioners allocate resources, develop programming for intervention, and inform supervision decisions. However, a good predictive performance is not enough and we also want the predicting factors to be interpretable, fair and based on meaningful behavioural proxies.

Current AI-based offender profiling systems sometimes make implicit attempts to deduce psychological or personal characteristics and frequently rely on intricate statistical correlations. These methods create moral and legal issues, especially when a prediction's justification is unclear. Furthermore, it is challenging to determine whether predictions are impacted by irrelevant or biased factors when models are opaque. These drawbacks underscore the necessity of frameworks that prioritise openness and concentrate on observable, behaviour-derived data rather than arbitrary or subjective characteristics.

In response to these issues we present a behaviour aware explainable AI framework for criminal offender profiling. We put forth the idea that which is to base recidivism predictions on identifiable behavioural trends from past criminal history. Also, instead of we try to determine psychological profiles we put forth behaviour-based features which include arrest frequency, custodial exposure, early involvement in crime, prison history, and offense severity which reflect what the criminal does with the justice system and also are easier for human decision makers to grasp and work with.

Our proposed framework combines ensemble machine learning models with cutting-edge explainability methods. In particular, we used SHapley Additive exPlanations (SHAP) to explain feature importance on both global and individual level as well du Local Interpretable Model-Agnostic Explanations (LIME) to provide instance-based explanations for each prediction of an offender. This fusion can support an analysis from different perspectives and a comparison of the global impact with that in individual cases.

The study used the public COMPAS data set which is a large element in recidivism prediction research. We report that which put forth model does well in terms of performance which also at the same time we see does very well in terms of interpretability. Also, we see that what we put forth in terms of which behavioural factors play key roles in recidivism like arrest record and early criminal involvement do in fact play large roles in what the models produce. Also, we note that the SHAP and LIME explanations we got out of the model are very much in line with what is found in the criminology research which in turn supports the model's validity.

2. Proposed Methodology

The approach seeks to guarantee that recidivism predictions can be made not only accurate but also transparent, interpretable, and anchored in observable behavioural phenomena. The general process involves data collection, the construction of behaviour-oriented features, predictive modelling, and explanation-driven interpretation.

2.1 Overview of the Proposed Framework

The proposed framework follows a structured pipeline comprising four major stages:

a) Dataset acquisition and pre-processing
b) Behaviour-aware feature engineering
c) Recidivism risk prediction using machine learning
d) Model explainability using XAI techniques

Unlike traditional black-box offender profiling systems, the framework straightforwardly behavioural factors motivated from previous criminal records. These indicators are subsequently analysed using explainable artificial intelligence techniques to ensure transparency at both global and individual levels.

2.2 Dataset Description

a) COMPAS Dataset Overview

Our proposed framework combines ensemble machine learning models with cutting-edge explainability methods. In particular, we used SHapley Additive exPlanations (SHAP) to explain feature importance on both global and individual level as well du Local Interpretable Model-Agnostic Explanations (LIME) to provide instance-based explanations for each prediction of an offender. This fusion can support an analysis from different perspectives and a comparison of the global impact with that in individual cases.

b) Database Structure

The dataset is provided in the form of a relational database containing multiple interlinked tables. The primary tables used in this study include:

- **people**: demographic and individual-level offender information
- **casearrest**: arrest records associated with criminal cases
- **jailhistory**: details of jail stay and custodial exposure
- **prisonhistory**: information related to prison terms
- **charge**: offense and charge-related information
- **compas** and **summary**: recidivism outcomes and assessment summaries

These tables are merged using common identifiers to construct a unified offender-level dataset suitable for behavioural analysis and modelling.

2.3 Data Pre-processing

Prior to modelling, several pre-processing steps are applied to ensure data consistency and quality:

- Removal of incomplete or inconsistent records
- Aggregation of multiple criminal records at the individual level
- Normalization and encoding of selected variables
- Handling of missing values using appropriate strategies

The pre-processing stage ensures that the dataset accurately reflects behavioural histories while avoiding redundancy and noise.

2.4 Behaviour-Aware Feature Engineering

A key contribution of this work lies in the construction of **behaviour-oriented features** that capture observable patterns of criminal activity. These features are derived from historical records and avoid psychological or subjective inference.

The following features are used in the proposed framework:

- **Arrest Count**: Total number of prior arrests associated with an offender, representing the frequency of criminal involvement.
- **Custody Intensity**: A measure reflecting repeated or prolonged custodial exposure, derived from jail history.
- **Early Criminal Onset**: A binary indicator representing whether the offender entered the criminal justice system at an early stage of life.
- **Prison Terms**: The number of prison sentences served by the offender.
- **Felony Ratio**: The proportion of felony-level offenses relative to total offenses, indicating offense severity.

These features are selected due to their interpretability and relevance to established criminological theories related to recidivism.

2.5 Predictive Modelling

After feature construction, the dataset is split into train and test subsets. An ensemble-based machine learning model is used to assess the risk of recidivism. The ensemble method is selected for its strength and capability to represent complicated and non-linear patterns in the behavioural data. The model outputs a yes/no prediction on whether the offender is likely to reoffend. The predictive performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05680

2.6 Explainability Layer

To address transparency concerns, an explainability layer is integrated into the framework using two complementary XAI techniques:

Global and Local Explanation using SHAP

SHapley Additive exPlanations (SHAP) are applied to quantify the contribution of each behavioural feature to the model's predictions. SHAP provides:

- **Global explanations**, identifying the most influential features across the entire dataset
- **Local explanations**, illustrating how individual features contribute to specific offender predictions

These explanations enable stakeholders to understand both overall trends and case-specific decision rationales.

Local Surrogate Explanation using LIME

Local Interpretable Model Independent Explanations (LIME) we use to put forth easy to understand instance level explanations. LIME puts forward a interpretable surrogate model which is a local approximation of the original model and also brings out the features that support or which in turn oppose a given prediction. We use LIME as a complement to SHAP which it does so by offering a different yet very readable explanation which in turn increases the models' transparency and our confidence in its decisions.

3. Results and Analysis

The analysis focuses on both **predictive performance** and **model interpretability**, with special emphasis on **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-agnostic Explanations)** to ensure transparency in recidivism risk prediction.

3.1 Model Prediction Overview

The model is calibrated to estimate the likelihood of recidivism versus non-recidivism using features extracted from arrest, custody and incarceration records characterizing behavioural and criminal history. Prediction results show that behavioural signals derived from criminal records over time substantially condition the model decision boundary. To move beyond black-box predictions, explainable AI techniques were employed to analyse:

- **Global feature importance** (SHAP summary plot)
- **Instance-level decision reasoning** (SHAP force plot and LIME explanation)

3.2 Global Feature Importance Analysis Using SHAP

Figure 1 presents the **SHAP summary plot**, which visualizes the overall contribution of each feature toward the model's recidivism predictions across the entire test dataset.

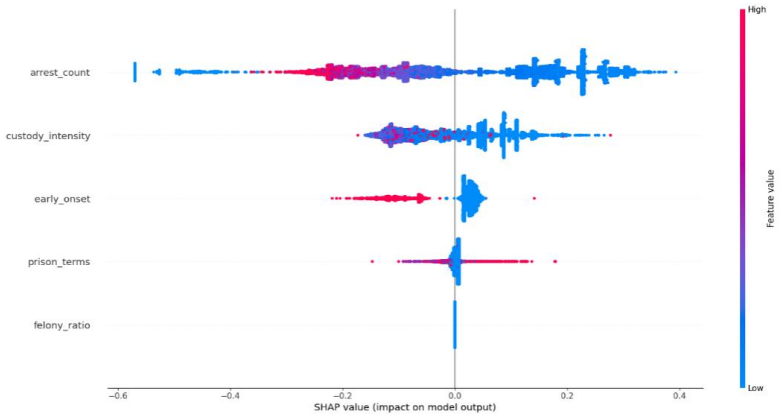

Figure 1. SHAP summary plot

Key Observations from Figure 1 (SHAP Summary Plot)

a) Arrest Count

- There is one, single feature which stands out as being the most influential.
- The larger the arrest count (shown in red), the more it dominates and leads to a higher recidivism risk, driving the model prediction upwards.
- Fewer arrest counts (blue) lead to a non-recidivism prediction.

b) Custody Intensity

- Is closely positively related to recidivism
- Those who have more custodial supervision will usually have a higher level of predicted risk.
- This suggests an attitudinal effect of cumulative or sustained contact with the justice system.

c) Early Onset of Criminal Behaviour

- Early entry involvement in crime provides much stronger estimates of the probability of recidivism.
- The model collection includes a theoretical assumption of criminology that earlier offenders are more prone to recidivation.

d) Prison Terms

- Less of an influence than number of arrests and level of custody.
- More time in prison increases the odds of recidivating slightly, and the variation decreases.

e) Felony Ratio

- Shows very low influence compared to other features.
- Suggests that frequency and timing of criminal behavior are better predictors than severity of offense alone.

Overall Insight:

SHAP analysis also verifies that behavioural history features play dominant roles on the prediction results, which supports our behaviour-informed design of the proposed framework.

3.3 Instance-Level Explanation via SHAP Force Plot

To provide an instance level explanation for the model decision, we use a popular visualization tool named SHAP force plot (Lundberg and Lee, 2017). SHAP Figure 2 shows a SHAP force plot for an offender describing how the values of its features affected the outcome. Explanation for Figure 2 (SHAP Force Plot) For the first data-point, from figure 1 discriminate_key (sa16): wherever the value in the first line is 0, that perturbation either remains mostly agnostic or has a negative effect. The person chosen has a high estimated probability of re-offense from the model (~0.90). Some of the features pushing toward recidivism prediction (red) are:

- arrest_count = 1.0
- custody_intensity = 0.0
- prison_terms = 0.0

The feature early_onset = 1.0 provides some correction to prediction, but not enough in order to effectively countersign the effect of cumulative arrest history. This figure vividly illustrates how various behavioural signals combine together to arrive at a high-risk rating.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05680

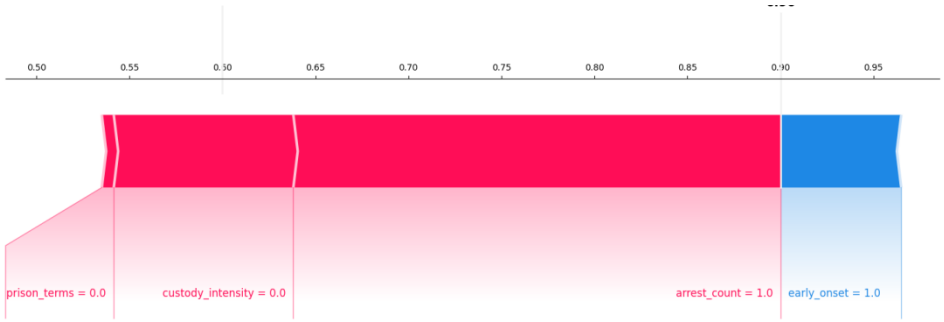

Figure 2. SHAP force plot

3.4 Local Explanation Using LIME

As a counterpart to SHAP, LIME was used to produce local interpretations for each prediction.
The LIME output indicates that:

- Arrest count ≤ threshold is the most influential local contributor of the recidivism class.
- Early initiation of criminal activity and contact with the justice system provide additional support for the prognosis.
- The definition is equivalent to SHAP results, which guarantees consistency of explanations across explainability options.

Significance: Consistency between SHAP and LIME provides increased confidence over the reasoning of the model and reduces concerns about instability of explanations.

3.5 Interpreting the Behaviour and Criminological Significance

From a behavioural criminology perspective:

- Repeated arrests are reflective of consistent antisocial behavioural trends.
- Custody intensity is a measure of system penetration, which can deepen criminal identification.
- Another widely-recognized indicator of chronic offending is an early age of onset.

The explainable AI framework enables theories like these to be effectively translated into measurable and interpretable model reasoning.

3.6 Summary of Findings

- Arrest history is the strongest risk factor for new recidivism.
- Dynamic behavioural trajectory characteristics are superior to static offense severity indicators.
- Explainability tools that expose transparent, consistent, and policy-relevant decision logic.
- The proposed framework balances **predictive accuracy with ethical accountability**, making it suitable for real-world criminal justice applications.

4. Conclusion and Future work

This paper reports a study that put forth a behaviour aware explainable AI framework for criminal offender profiling which we have a focus on transparent and interpretable recidivism prediction. We integrated behavioural variables like arrest count, custody intensity, and early criminal onset into machine learning models which in turn produced effective risk predictions also it broke away from black box decision making. We used SHAP and LIME which in turn gave us global and instance level explanations which in what which also report how behavioural elements play a role in model results. We found that the results of the study to be consistent with criminological theories which in turn which improved the trust, accountability and practical application in criminal justice decision support.

Future study may look into improving this framework by which we introduce temporal and sequential elements that in turn will better represent long term behavioural change also as we incorporate fairer methods of assessment which in turn will reduce bias across different demographic groups. Also, we will include socio economic and rehabilitation related variables which in turn will improve prediction accuracy and intervention relevant. Also, we will put to test the framework in many different jurisdictions and put it out as a real time decision support which in turn will improve its broad-based acceptance and real-world application.

Reference

1. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Lulu Press.
2. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607
3. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
4. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
5. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).
8. Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44. https://doi.org/10.1177/0049124118782533
9. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047
10. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
11. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324