Interpretable Machine Learning to Identify Risk Factors for Recidivism in Intimate Partner Violence

Çerağ Oğuztüzün, BS[1], Mehmet Koyutürk, PhD[2], Günnur Karakurt†, PhD[1]
†Corresponding author

[1]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106, USA; [2]Department of Psychiatry, Case Western Reserve University, University Hospitals Cleveland Medical Center, Cleveland, OH 44106, USA

Abstract

Intimate Partner Violence (IPV) remains a significant global health issue with severe consequences ranging from physical injury to death, with rates rising in recent years. Prediction of recidivism is critical for prevention and treatment. Using data from a four-year clinical study, we develop interpretable machine-learning models to identify features for physical assault recidivism among IPV offenders. To standardize clinician-assigned severity scores and address non-linear associations, we apply filtered target encoding, which reduces subjectivity and bias in assessment. We find that combining self-reported and partner-reported variables enhances predictive power. Through feature importance analyses, we identify factors associated with lower recidivism risk, including decreased substance use and avoiding partner contact, while separation processes correlate with higher reoffending likelihood. These findings advance IPV risk assessment by providing a deeper understanding of risk factors critical for improving treatment effectiveness and addressing disparities in IPV management.

1 Introduction

Intimate partner violence (IPV) is physical, sexual, or emotional abuse by a current or former partner [1], and affects roughly one in four women globally, carrying substantial morbidity and mortality. IPV disproportionately impacts women and marginalized groups, driven by systemic inequities, limited access to quality care, and adverse childhood experiences [2, 3, 4].

IPV's adverse effects span acute physical injuries (bruises, lacerations, fractures), disabilities (e.g., vision or hearing damage), chronic conditions (headaches, chronic pain), systemic ailments (asthma exacerbations, hypertension, irritable bowel syndrome), and mental-health disorders (depression, PTSD, substance use, suicidality) [5, 6, 7, 8, 9]. Predicting who will reoffend is critical for reducing repeat harm, guiding scarce clinical resources, and informing policy. Recidivism rates for IPV range from 20–35% even after intervention, perpetuating cycles of injury and psychological trauma for survivors [10]. Early identification of high-risk individuals enables targeted interventions (counseling, case management) and multi-agency responses shown to lower reoffense rates and improve safety outcomes [11].

Existing studies laid important groundwork: Pham et al. (2023) [12] showed ODARA and SARA yield modest discrimination (AUC ≈ 0.59); Yu et al. (2023) [13] reported registry-based Cox models on 27,456 Swedish DV arrests achieving AUC ≈ 0.75 for general reoffending but only 0.63 for DV-specific rearrest; and Collins et al. (2021) [14] found prior DV arrests and longer case-processing times predict higher recidivism. Our approach builds on these by transforming interviewer-rated severity into dynamic, case-specific risk signals rather than relying on fixed or population-level scores.

The Minneapolis Domestic Violence Experiment found initial benefits of mandatory arrest in reducing reoffending [15], but later studies observed attenuated effects [16, 17], spurring more counseling-based and personalized interventions. A variety of court-mandated and voluntary programs (from Duluth sociocultural models to cognitive-behavioral and group therapies) have shown mixed results [18, 19, 20, 21, 22].

Despite numerous established risk indicators, existing assessment tools often perform little better than chance [21, 23]. Exploratory models can fill this gap by detecting treatment lapses early [24] and enabling tailored, multi-agency interventions before violence escalates [25].

In this study, we enhance insights from predictive models of IPV recidivism. We first benchmark multiple machine learning models and analyze feature importances, noting prevalent non-linear relationships in subjective scores. To

Table 1: **List of the variables used in the predictive models**, sorted with respect to feature importance (according to our results). The linearity of the relationship of each ordinal or continuous variable with the outcome variable, according to our results, is shown in the last column. Please see Methods for the description of the procedure for determining the linearity or non-linearity of the variables

Variable	Definition	Data Type	Linearity
O1	**Outcome Variable:** Was there a re-assault incident that occurred between the 3rd and 15th month of follow-up?	Binary	–
P3SEVX	Have you ever committed a severe assault, ranging from kicking to sexual assault?	Binary	–
S3TMP	Have you been tempered in the past 3 months?	Binary	–
P3TOTX	What is the frequency with which your partner has taken actions toward you, whether in self-defense, fear, revenge, or for any other reason?	Ordinal	Linear
B3TOTX2	In the past 3 months, how many times have you tried to control or engage in aggressive behaviors toward your partner, such as limiting her social interactions, affecting her finances, or causing verbal or physical harm?	Ordinal	Linear
P3MINX	Have you ever engaged in minor forms of physical aggression, such as pushing or slapping?	Binary	–
ATKWN2	Has any partner you've been involved with physically attacked you in the last 3 months?	Binary	–
B3VRBX2	In the past 3 months, have you engaged in any verbal aggression, property damage, or harm to pets towards your partner?	Ordinal	Linear
DRUN	How often were you drunk or high?	Ordinal	Linear
DHPRT	How often was your partner drunk or high?	Ordinal	Linear
DHPRTX	How often was your partner drunk or high? *(Reported by the partner)*	Ordinal	Non-linear
RPPLC	Has any partner ever sought shelter, legal action, or professional help as a result of your behavior towards her?	Binary	–
P3TKN	Has your partner ever threatened you with a knife?	Binary	–
DOFT	How often do you drink alcohol or use drugs?	Ordinal	Non-linear
FRSTMX	When was the last time your partner started severely physically assaulting you? *(Reported by the partner)*	Binary	–
TMINV	How long have you been intimately involved with your current partner?	Continuous	Non-linear
DDFRQ2	Are you using alcohol or drugs more or less in the past 3 months?	Ordinal	Non-linear
SELIVA2	What is the frequency of your encounters with the partner you've abused in the past three months?	Ordinal	Non-linear
MARSTX	What is your current marital status?	Categorical	–
DDPRT	How often does your partner drink alcohol or use drugs?	Ordinal	Non-linear

standardize interviewer-assigned severity and mitigate bias, we develop a *filtered target encoding* technique that preserves ordinal information while smoothing inconsistencies. Our key contributions are:

- A novel encoding method that refines ordinal variables based on training-outcome statistics and adjacent-category smoothing, improving model interpretability and reliability.
- Comprehensive feature-importance analyses and visualizations of risk-factor relationships with recidivism.
- Demonstration that integrating self- and partner-reported variables enhances understanding of recidivism predictors.

2 Methods

Data Source

Batterer intervention programs target men arrested for domestic violence. Here, we utilize data from four different batterer intervention programs, spanning 840 *participants* and their female *partners*, including initial and subsequent partners [26]. Data was collected longitudinally from treatment programs in Denver, Pittsburgh, Dallas, and Houston.

At the beginning of the program, demographic details and baseline measures of personality and alcoholism were gathered using the Millon Clinical Multiaxial Inventory, Version III (MCMI-III), and the Michigan Alcoholism Screening Test (MAST). Severe psychopathology was indicated by scores above 74 on any of the six MCMI clinical subscales, and a score above 4 on the MAST suggested likely alcoholism. Additional intake data included demographic characteristics, employment status, education level, substance use, living situation, and history of violence. Follow-up interviews were conducted by phone at three-month intervals about re-assault as well as other forms of abuse, partner contact, substance use and abuse, further arrests, and the use of additional services and treatment. Complete data from batterers, their initial partners, and any new partners available for 65% of the sample over 15 months. For more details, please see Gondolf et al.'s study [18]. A subset of 109 men and their female partners from the initial pool remained active in the study, with regular quarterly interviews being conducted, yielding 32 recidivism-positive cases and 77 recidivism-negative cases. These responses were further supplemented by a thorough examination of arrest records, and new variables were created to monitor the accounts provided by the participant and their partner [27].

An earlier examination of this data indicated that men from Sites II-IV, despite variations in their program structures, showed no significant differences in recidivism rates. In contrast, some differences were noted at Site I, as documented in the study by Gondolf [28]. Taken together, these observations indicate significant heterogeneity between and within treatment programs, suggesting that interpretable machine learning can inform personalized treatment of batterers.

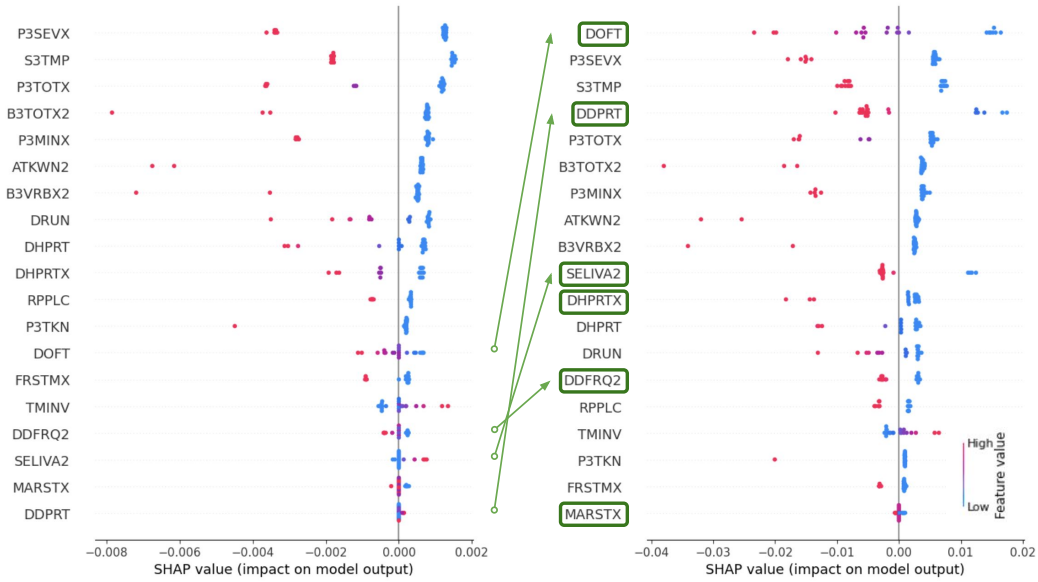


Figure 1: **Feature importance analysis using Shapley Additive Explanations (SHAP) before and after encoding features.** For each feature, the SHAP value assigned to the feature for each instance of the support vector machine (SVM) model is shown. Features are ordered from most important to least important. The red color signifies the feature value being high and the blue color signifies the low feature value for positive samples. The model before encoding is shown on the right, and the model after encoding is shown on the right. Encoded variables are highlighted with a green frame. Arrows indicate features that have risen in the order of feature importance after encoding.

In the following discussion, we refer to the men who are enrolled in these programs (and represent one instance in our data) as *the participant* and their female partners who are subjected to violence as *the partner*.

Variables

In our experimental setting, we use the background survey data. The survey consists of various questions assessing the nature and frequency of aggressive behaviors, substance use, consequences of behavior, and relationship dynamics among participants. We represent the responses to survey answers as binary, categorical, ordinal, or continuous variables as applicable (Table 1). We define the outcome variable as physical re-assault between 3-15 months follow-up.

Predictive Models

We train a range of traditional machine learning algorithms to predict recidivism using all available features, including SVM, logistic regression, Naïve Bayes, K-nearest neighbor, Random Forest, XGBoost, AdaBoost and CatBoost. Since the number of samples is relatively low, we do not consider deep learning. We evaluate the performance of the resulting models using 5-fold cross-validation and observe that the models exhibit modest accuracy in predicting recidivism, with accuracy plateauing around 60% for most algorithms. While this can be attributed to the relatively small sample size, delineation of the complexity of the relationship between recidivism and other variables may provide insights in the prediction of recidivism and may inform decision-making in treatment. Motivated by this consideration, we analyze feature importance in detail to elucidate this relationship.

We use SHAP [29] to assess feature importance in the SVM model and focus on the top 20 most important features (Fig. 1). We observe that several variables exhibit non-linear association with recidivism. For example, the SHAP values of *Partner-reported Frequency of Alcohol or Drug Use* (DHPRTX) are clustered into three groups, with a large purple group concentrated close to 0. To investigate the reason for this behavior, we assess the distribution of this variable for instances with vs. without recidivism (Fig. 5, Row 3, Column 1). As seen in the figure, the recidivism-negative group shows a consistently decreasing pattern (53%, 32%, 15%), while the recidivism positive group exhibits a U-shaped distribution (44%, 25%, 31%) with category 3 ('Often/Very often') showing higher values than category 2 ('Sometimes'). While this non-linear pattern makes it challenging for a predictive model to effectively utilize DHPRTX to predict recidivism, the association is evident in the distributions.

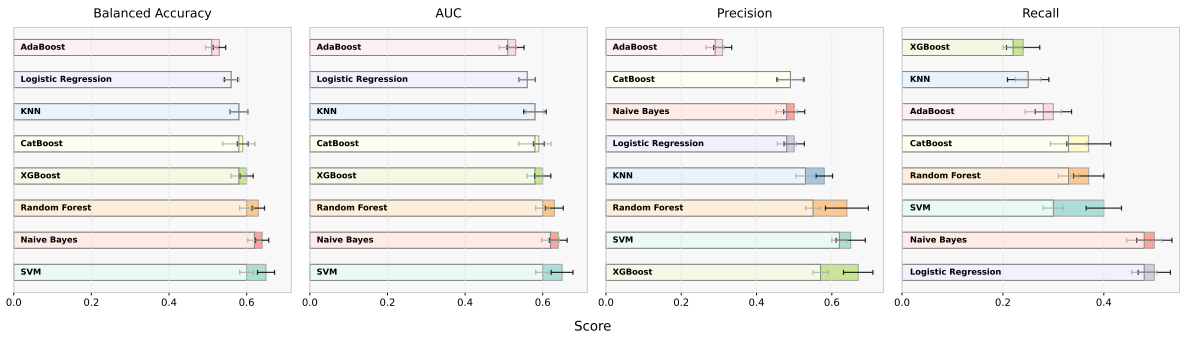


Figure 2: **Comparison of classification algorithm performance in predicting recidivism.** Performance metrics (Balanced Accuracy, AUC, Precision, and Recall) are shown for eight classification algorithms using filtered target encoding (darker tops, window size $W = 3$) versus no encoding (lighter bottoms). Error bars represent 95% confidence intervals from 10-fold cross-validation.

Filtered Target Encoding

To improve the ability of machine learning algorithms to capture non-linear associations, we develop a *Filtered Target Encoding* technique to transform variables. Encoding of variables is particularly suited to our application and the nature of the data since most variables are coded based on the subjective interpretation of participants and interviewers. In machine learning, Target Encoding [30] uses the training instances to replace the categories in a variable with the means of the outcome variables for those categories. This helps the classifier resolve the complex relationships between each variable and the outcome. An additional benefit Target Encoding directly offers is the ability to interpret the association between each individual variable and the outcome.

As seen in Table 1, most of the variables that require transformation (marked as non-linear) are discrete ordinal variables that are coded somewhat subjectively. For example, the variable DOFT (*Alcohol or Drug Consumption Frequency*) takes integer values from 1 to 9, where 1 indicates never drinking and 9 indicates drinking everyday. Clinicians often use ordinal variables to assess physical or mental health. However, these evaluations can vary between clinicians; one practitioner may assign a '1' for a variable, whereas another may assign a '2' for the same symptom severity. Similarly, a response of "2" to this question may indicate slightly different frequencies of alcohol or drug consumption. The reliability and validity of downstream statistical models can be impacted by such inconsistencies, introducing noise and bias into the analysis.

We address this challenge by developing a novel encoding method derived from target embeddings of ordinal variables. Our technique is based on the premise that *closer values of such ordinal variables indicate similar levels* as compared to more distant values. Thus by applying a filter function over a window, i.e., by calculating a weighted ratio of outcomes associated with each ordinal value, our technique aims to "standardize" these ordinal values. Letting x_i denote a specific value of variable X, we use the training instances to compute the encoding for x_i as follows:

$$\text{Encoding}(x_i) = \frac{\sum_{j=-w}^{w} \alpha_j \times \text{Count}(X = x_{i+j}, Y = 1)}{\sum_{j=-w}^{w} \alpha_j \times \text{Count}(X = x_{i+j})} \tag{1}$$

In this formulation:

- x_i denotes a specific value of variable X (For example, if X is DOFT, x_i could be any integer between 1 and 9)
- $\text{Count}(X = x_{i+j}, Y = 1)$ denotes the number of training instances where $X = x_{i+j}$ and $Y = 1$, where Y is the binary outcome variable
- $\text{Count}(X = x_{i+j})$ denotes the number of all training instances where $X = x_{i+j}$
- $w = (W - 1)/2$ where W denotes the window size (an odd number), representing the range of values in the neighborhood of x_i that are considered when computing the encoding
- α_j denotes the filter function weights, where:
  - $\alpha_0 = 2$ (giving double weight to the exact value x_i)
  - $\alpha_j = 1$ if $0 < |j| \leq w$ (equal weight to neighbors within window), $\alpha_j = 0$ otherwise (zero weight to values outside window)

A step-by-step example of the formulation is presented in the Appendix. The encoding can be tuned by adjusting the window size W and the filter function α. In our computational experiments, we consider different values of W ranging from 1 to 5 and assess the predictive performance of the machine learning models as a function of W (Fig. 3). It is important to note that while this filtering approach with window size $W > 1$ is applied to ordinal variables (where neighboring values have semantic relationship), categorical variables such as MARSTX (*Marital Status*) use a window size of $W = 1$. This is because categories in purely categorical variables (as opposed to ordinal categories) have no inherent ordering or proximity relationship that would make averaging across neighboring categories meaningful. In other words, there is no logical "neighboring" relationship between being single, married, or divorced that would justify applying the filtering window.

Once an encoding for each variable is computed, we replace the values of the variables for each training instance with the encoded value and train the machine learning model on the encoded training data. When a test instance comes in, the values of variables are similarly transformed using the encoding function, and the model is applied to the encoded instance to predict the outcome for that instance. This strategy avoids data leakage as the test instances are never considered while computing the encoding and no normalization or any other pre-processing is applied to transform the values of training instances alongside test instances.

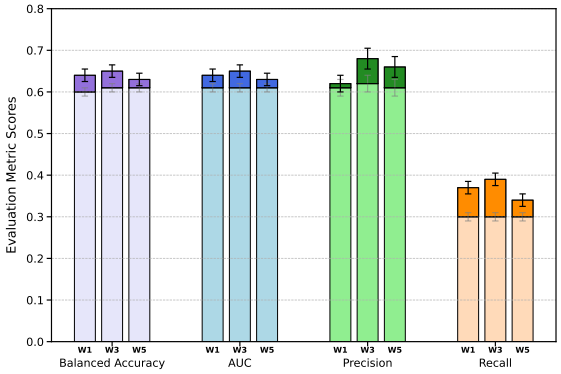


Figure 3: **Filtering and window size effects on IPV recidivism prediction.** Bars display SVM performance comparing filtered features (darker tops) across window sizes ($W = 1, 3, 5$) vs. original features (lighter bottoms) using four performance criteria. Error bars show 95% confidence intervals from 10-fold cross-validation.

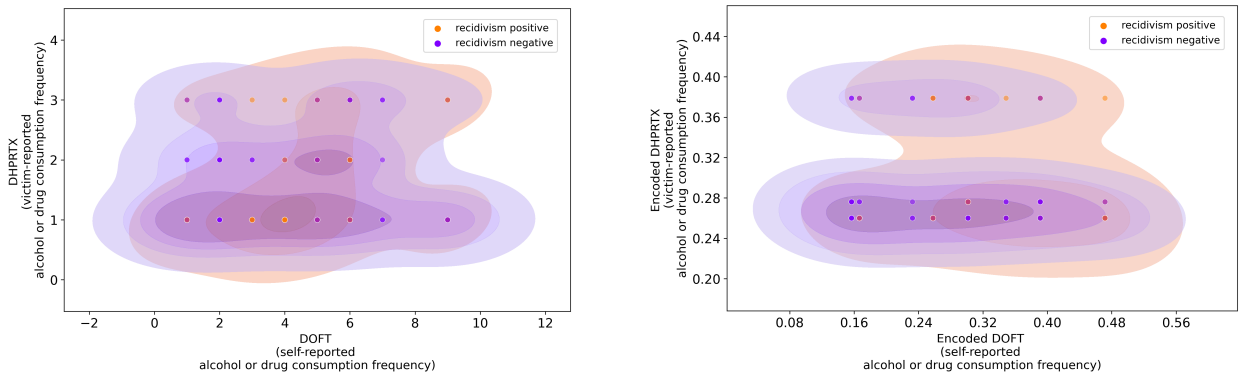
3 Results

We assess the effectiveness of the proposed target encoding technique using a range of traditional and state-of-the-art classification algorithms (Fig. 2). Notably, conventional algorithms like SVM and Naive Bayes benefit more from target encoding as compared to contemporary methods like XGBoost. However, more sophisticated models do not outperform SVM and Naïve Bayes when filtered features are used. This could be because of the limited size of the training data, suggesting that more sophisticated models may benefit from more extensive training data.

When we examine the effect of encoding and window size on SVM model's performance (Fig. 3), it is evident that encoding non-linear variables enhances classification performance. Furthermore, the variance in the prediction performance of the classifier diminishes when non-linear variables are encoded, suggesting that encoding also enhances the reliability of the models. Notably, optimal performance is achieved with a window size of 3, indicating that filtering is indeed useful, but excessive smoothing of variables can hamper performance. Filtered target encoding enhances the predictive model's recall (with about 30% improvement for $W = 3$), suggesting that encoding can drastically improve a model's ability to catch cases with potential for recidivism. In addition to enhancing predictive model performance, filtered target encoding reveals relationships between recidivism and its predictors. Next, we present insights from our feature importance analyses using this encoding.

Participant's Frequency of Alcohol and Drug Usage. This variable is measured based on both participant's (DOFT) and partner's (DHPRTX) reports. Inspecting Fig. 1, we observe that infrequent users of alcohol or drugs are more distinctively characterized as less likely to re-offend, since the model emphasizes the negative correlation between low substance use and recidivism. The transformation of this variable makes its distribution clearly more separated for participants with or without recidivism become more clearly separated (Fig. 5, Row 4, Column 3)), emphasizing the impact of encoding techniques on enhancing the predictive ability of important features.

The encoding map of the participant-reported substance use variable (DOFT) in an inverted U-shape (Fig 5, Row 4, Column 2) reveals that there is a range of increased substance use that is most strongly associated with recidivism. In contrast, the encoding map of the partner's report (DHPRTX) is monotonically increasing (though not linear) - suggesting that the non-linear relationship of the participant-report variable with recidivism may be due to reporting



(a) Using Original Data

(b) Using Encoded Data

Figure 4: **Relationship between the participant's self-reported (DOFT) and partner-reported (DHPRTX) alcohol/drug consumption frequency and its association with recidivism.** Dots represent individual cases colored by recidivism status (orange = recidivism positive, purple = recidivism negative). Contour lines show the density distribution of cases for each recidivism group. **(a)** Using original data shows overlapping distributions with limited separation between recidivism groups. **(b)** Using target-encoded data reveals clearer patterns; partner reports (encoded DHPRTX) gain predictive importance when participants report moderate substance use (encoded DOFT).

inconsistency, e.g., high-substance-use participants may be under-reporting their use of alcohol or drugs. To investigate this further, we scatter-plot the DOFT and DHPRTX variables (Fig. 4) to see whether there are any inconsistencies between the partner-reported and self-reported alcohol or drug consumption frequency. While there is some inconsistency between reports provided by the participant and partner, inconsistency is not necessarily indicative of recidivism by itself. Based on this observation we conclude that these variables should be used together such that if the participant reports a high frequency of substance use, the partner's report gains importance, whereas if the participant reports a medium frequency of substance use, recidivism is more likely independent of the partners' report. This observation also highlights the need for more biologically reliable measurements of drug and alcohol usage.

Partner's Frequency of Alcohol and Drug Usage (DDPRT). The SHAP values for this variable are near zero before encoding, suggesting that the model considered partners' substance use frequency somewhat neutral. However, there is distinct separation in SHAP values assigned to this variale after encoding, suggesting that recidivism has a stronger relationship with partner's substance use frequency. Inspection of the encoding map (Fig. 5, Row 2, Column 2) suggests that moderate frequency of substance use by the partner is associated with a decreased likelihood of recidivism as compared to rare or no substance use by the partner. However, the highest frequency of substance use by the partner are associated with the highest likelihood of recidivism.

Participant's Frequency of Encountering the Partner in the Past 3 Months (SELIVA2). Before encoding, the histograms of this variable for cases with and without recidivism have similar shapes, where most participants rarely encounter their partner, but there are representatives of all frequencies in both groups (5, Row 5, Column 1). After encoding, the relationship becomes more nuanced, where the distribution of the encoded variable is bi-modal for both recidivism and non-recidivism, and the left-side peak for recidivism almost diminishes (5, Row 5, Column 3). Inspection of the encoding map (5, Row 5, Column 2) suggests that this is due to recidivism being associated with either very low frequency of encounters or high-frequency encounters. The encoding map also suggests that highest frequency of encounters between the participant and the partner are linked to a reduced likelihood of recidivism.

Participant's Alcohol and Drug Consumption Frequency over The Past 3 Months (DDFRQ2). The SHAP values for this variable between -0.01 and 0.01 (Fig. 1) show that an individual's recent substance use frqueency is marginally informative on potential recidivism. However, when we consider the encoded variable (Fig. 5, Row 1), we observe a monotonic but non-linear relationship between recent substance use and recidivism, where the more the recent substance use, its association with recidivism goes up.

4 Discussion

The key objective of this study is to understand the factors that contribute to recurrent IPV. Improving batterer treatment programs requires a deeper understanding of risk factors for future violence. While numerous risk indicators have been identified, existing risk assessment tools often struggle to accurately predict future violence, often performing marginally better than chance. To address this limitation, it is essential to examine the dynamic progression of violence within intimate relationships and individuals' responses to treatment by reducing bias that stems from the subjective coding of variables. Developing models that capture these patterns can help bridge a critical gap by enabling early identification of individuals at risk of reoffending. We also aim to enhance the accuracy and interpretability of the predictive models' by transforming the coding of target variables that are reported by participants and interviewers.

The encoded values are calculated to directly reflect the distribution of the outcome variable, which is recidivism, within the context of the original feature values. The encoding process, especially with a method like filtered target encoding, can capture non-linear relationships between the feature and outcome. This is because it considers the context of neighboring values, or the "window," when determining the encoded value. This can reveal patterns that a simple linear encoding might miss. Additionally, we can observe how encoding enhances the discriminatory value of the variables based on the separability of two peaks in the density plots (Fig. 5 - Encoded Distribution Column).

Our findings suggest that substance use patterns should be assessed in combination, considering that self-report and partner-report are two of the most critical factors in predicting recidivism (Fig. 4). Specifically, when participants report high frequency of substance use, their partner's account becomes more influential in predicting recidivism. In contrast, moderate substance use by the participant is associated with a higher likelihood of recidivism, regardless of the partner's report. This may be due to a potential association between truthfulness in reporting and recidivism, e.g., those who are likely to reoffend may be under-reporting their own substance use. Since it would be difficult to assess truthfulness in reporting, this observation suggests that effective integration of self-report and partner-report data is essential in assessing the likelihood of reoffending.

Moderate substance use by the offender's partner appears to be associated with a lower risk of recidivism compared to rare or no substance use. Recent substance use is increasingly associated with recidivism, suggesting a dose-dependent relationship between substance use patterns and the risk of repeat IPV. The amount consumed is possibly not clear, emphasizing the need for more objective, biologically reliable measures of drug and alcohol consumption. However, many biomarkers of substance use issues are not specific enough to be used reliably [31] . Intermittent or occasional alcohol consumption may exert little effect on peripheral inflammatory markers, while chronic and frequent alcohol use is associated with more pronounced clinical and biological consequences, similar to neuroimaging biomarkers [32].

By recognizing key progression thresholds early, these models allow for timely intervention adjustments, potentially preventing escalation into more severe violence. Incorporating substance abuse treatment into IPV interventions yielded better results compared to programs that did not have these components [20]. Past research also found that motivational enhancement techniques improve preparedness for change and reduce resistance and high dropout rates during the treatment process [33, 34]. Taken together, our results suggest that dynamic modeling of the interplay between relationship factors, risk factors of conflict and violence in the relationship, and combined data interpretation during treatment hold the key to accurately predicting treatment effectiveness and guiding decisions. Although our method outperforms traditional tools, accuracy remains modest ($\approx$60%), underscoring the inherent difficulty of predicting IPV recidivism. Factors such as unobserved or underreported psychological, relational, and contextual variables, the evolving nature of violent behavior, and unreliable self- or partner-reports all add noise. Moreover, with only 32 recidivism-positive cases, our estimates may lack generalizability and have high variance.

ACKNOWLEDGMENTS

This publication was made possible by US National Health Institutes (NIH) grant R01-LM012518 from the National Library of Medicine. We thank the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan for providing data access. The findings and conclusions in this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or ICPSR. Correspondence

concerning this article should be addressed to Gunnur Karakurt, Department of Psychiatry, Case Western Reserve University. Cleveland, OH 44109. E-mail: gkk6@case.edu

References

1. Breiding M, Basile KC, Smith SG, Black MC, Mahendra RR. Intimate partner violence surveillance: uniform definitions and recommended data elements. Version 2.0. 2015.
2. Rosay AB. Violence against American Indian and Alaska Native women and men. 2016.
3. Fahmy E, Williamson E, Pantazis C. Evidence and policy review: Domestic violence and poverty. York: Joseph Rowntree Foundation. 2016.
4. Oram S, et al. The Lancet Psychiatry Commission on intimate partner violence and mental health: Advancing mental health services, research, and policy. The Lancet Psychiatry. 2022;9(6):487-524.
5. Smith SG, Zhang X, Basile KC, Merrick MT, Wang J, Kresnow Mj, et al.. The national intimate partner and sexual violence survey: 2015 data brief–updated release; 2018.
6. Karakurt G, Patel V, Whiting K, Koyutürk M. Mining electronic health records data: Domestic violence and adverse health effects. Journal of family violence. 2017;32(1):79-87.
7. Garcia-Moreno, et al. Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. WHO; 2013.
8. Organization WH, et al. Understanding and addressing violence against women: Intimate partner violence. World Health Organization; 2012.
9. Karakurt G, Yılmaz S, Kumari M, Gao K, Koyutürk M. Comprehensive analysis of electronic health records to characterize the association between intimate partner violence and mental health. AMIA Summits on Translational Science Proceedings. 2023;2023:310.
10. Verdugo-Martínez A, Ronzón-Tirado R, Redondo-Rodríguez N. Personality traits and their role in intimate partner violence recidivism: A 15-year follow-up study within a prison sample. Personality and Individual Differences. 2025;235:112969.
11. Council NR, et al. Can interventions reduce recidivism and revictimization following adult intimate partner violence incidents. In: The Evidence for Violence Prevention Across the Lifespan and Around the World: Workshop Summary. National Academies Press (US); 2014. .
12. Pham AT, Hilton NZ, Ennis L, Nunes KL, Jung S. Predicting recidivism in a high-risk sample of intimate partner violent men referred for police threat assessment. Criminal Justice and Behavior. 2023;50(5):648-65.
13. Yu R, Molero Y, Lichtenstein P, Larsson H, Prescott-Mayling L, Howard LM, et al. Development and validation of a prediction tool for reoffending risk in domestic violence. JAMA network open. 2023;6(7):e2325494-4.
14. Collins AM, Bouffard LA, Wilkes N. Predicting recidivism among defendants in an expedited domestic violence court. Journal of interpersonal violence. 2021;36(13-14):NP6890-903.
15. Sherman LW, Berk RA. The specific deterrent effects of arrest for domestic assault. American sociological review. 1984:261-72.
16. Erez E, Belknap J. In their own words: Battered women's assessment of the criminal processing system's responses. Violence and victims. 1998;13(3):251-68.
17. Reid MR, Buchanan NT. Systemic biases promoting the under-inclusion of marginalized groups in randomized controlled trials for co-occurring alcohol use and posttraumatic stress disorder: an intersectional analysis. Journal of ethnicity in substance abuse. 2024:1-26.
18. Gondolf EW, et al. Batterer intervention systems: Issues, outcomes, and recommendations. Sage; 2002.
19. Pence E, Paymar M, Ritmeester T. Education groups for men who batter: The Duluth model. Springer; 1993.
20. Karakurt G, Koç E, Çetinsaya EE, Ayluçtarhan Z, Bolen S. Meta-analysis and systematic review for the treatment of perpetrators of intimate partner violence. Neuroscience & Biobehavioral Reviews. 2019;105:220-30.
21. Travers Á, et al. The effectiveness of interventions to prevent recidivism in perpetrators of intimate partner violence: A systematic review and meta-analysis. Clinical Psychology Review. 2021;84:101974.
22. Bell C, Coates D. The effectiveness of interventions for perpetrators of domestic and family violence: An overview of findings from reviews. Australia's National Research Organisation for Women's Safety; 2022.

23. Labarre M, Brodeur N, Roy V, Bousquet MA. Practitioners' views on IPV and its solutions: An integrative literature review. Trauma, Violence, & Abuse. 2019;20(5):679-92.
24. Khurana B, Seltzer SE, Kohane IS, Boland GW. Making the 'invisible'visible: transforming the detection of intimate partner violence. BMJ quality & safety. 2020;29(3):241-4.
25. Taccini F, et al. Perceptions of Intergenerational Transmission of Violence and Parenting Practices Among Pregnant Women and their Abusive Partners. Journal of Family Violence. 2024:1-17.
26. Heckert DA, Gondolf EW. Multi-Site Evaluation of Batterer Intervention in Pennsylvania, Texas, and Colorado, 1995-1999. 2015.
27. Ip EH, et al. Latent Markov model for analyzing temporal configuration for violence profiles and trajectories in a sample of batterers. Sociological Methods & Research. 2010 Nov;39(2):222-55.
28. Gondolf EW. Patterns of reassault in batterer programs. Violence and Victims. 1997 Jan;12(4):373-87.
29. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. NIPS; 2017. p. 4765-74.
30. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter. 2001;3(1):27-32.
31. Ruwel AG, Scherer JN, Silvello D, Kessler FHP, von Diemen L, Schuch JB. Hematological inflammatory biomarkers in patients with alcohol and cocaine use disorders. Trends in Psychiatry and Psychotherapy. 2024.
32. Ekhtiari H, Sangchooli A, Carmichael O, Moeller FG, O'Donnell P, Oquendo MA, et al. Neuroimaging biomarkers of addiction. Nature Mental Health. 2024:1-20.
33. Karakurt G, Koç E, Katta P, Jones N, Bolen SD. Treatments for female victims of intimate partner violence: systematic review and meta-analysis. Frontiers in psychology. 2022:44.
34. Oğuztüzün Ç, Koyutürk M, Karakurt G. Systematic investigation of meta-analysis data on treatment effectiveness for physical, psychological, and sexual intimate partner violence perpetration. Psychosocial intervention. 2023;32(2):59.

Appendix

Step-by-step example for target encoding DOFT DOFT is the self-reported "Alcohol or Drug Consumption Frequency" on a scale from 1 (never) to 9 (daily). Suppose at training time you see these counts for values near "3" on DOFT (occasionally):

- **Recidivism-positive cases:** DOFT = 2 ("rarely"): 5 people re-offended; 3 ("occasionally"): 12; 4 ("sometimes"): 7.
- **Total cases:** DOFT = 2: 20 participants; 3: 40; 4: 30.

1. **Why gather neighbors?** Raw DOFT "3" might come from different interviewers or subjective assessments (some "3"s are closer to "2" in risk, others closer to "4"). Pulling in counts at 2, 3, 4 smooths out that noise.
2. **We weight the center more and compute the weighted "risk" numerator:** We set $\alpha_0 = 2$ for DOFT=3 (in order to trust the exact "occasionally" value more) and $\alpha_{-1} = \alpha_{+1} = 1$ for its neighbors. We are counting "how many re-offenders" are around that level (while giving extra weight to the exact category): $1\times5+2\times12+1\times7 = 36$.
3. **We compute weighted total denominator:** We are counting "how many total people" are around that level: $1 \times 20 + 2 \times 40 + 1 \times 30 = 130$.
4. **We get the encoded value:** We can now interpret "3" not as the raw category but as a 27.7% "risk score" of recidivism:$\mathrm{Encoding}(3) = \frac{36}{130} \approx 0.277$.
5. **Model uses the risk score instead of "3":** In both training and testing, we replace the raw DOFT=3 with 0.277. That way, the classifier sees a smoother and probability-like signal instead of an arbitrary integer.

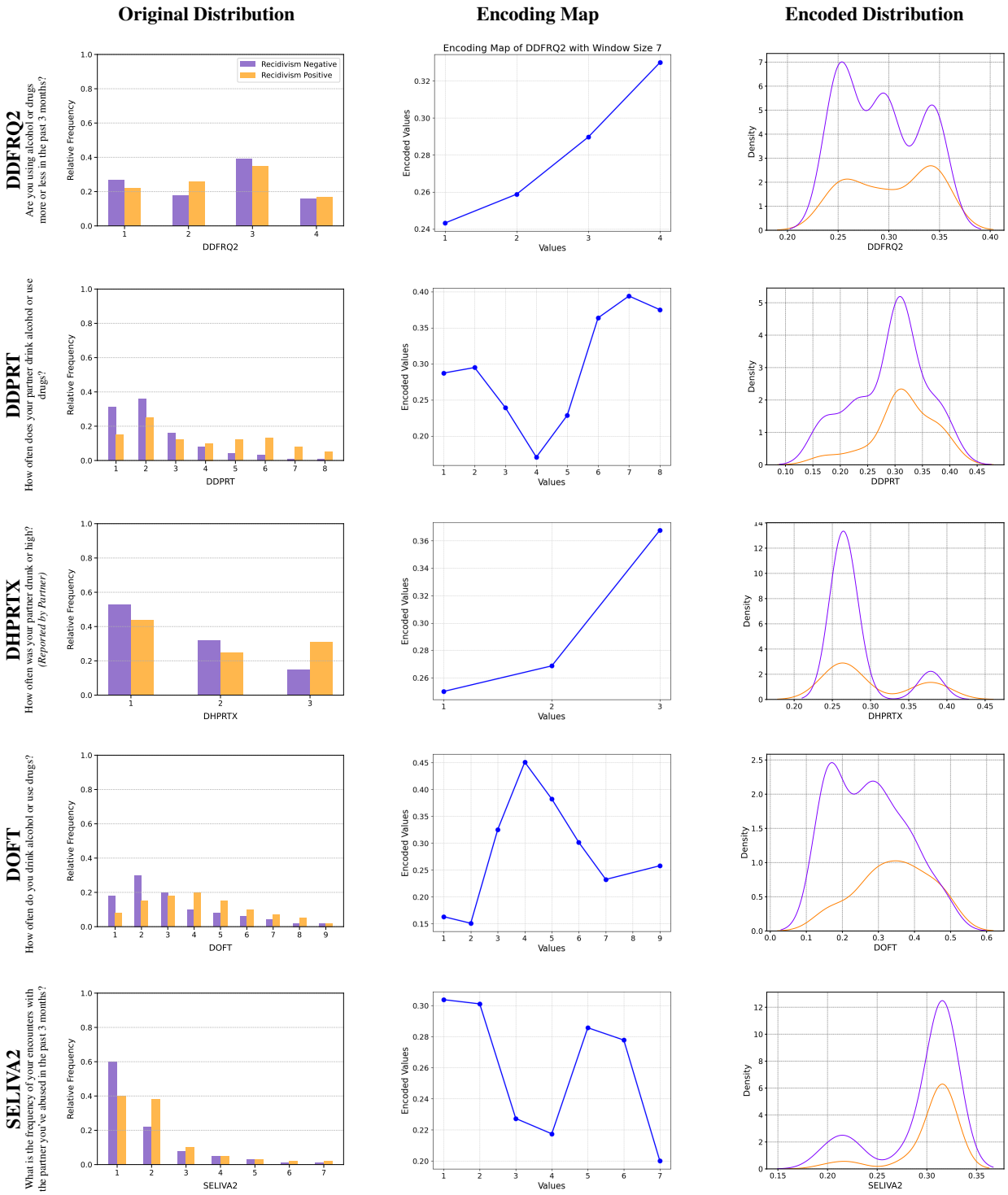


Figure 5: **The effect of filtered target encoding on select variables' association with recidivism.** Each row represents a variable. **Left:** Original distribution of the variable in cases with and without recidivism. **Middle:** Encoding map showing the correspondence between the original value and the encoded value for the variable, also indicating the likelihood of recidivism given the value (and its neighborhood, as the encoding is filtered). **Right:** Distribution of the encoded variable in cases with (in orange) and without (in purple) recidivism.