Open Access Library Journal
2026, Volume 13, e15141
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Machine Learning Prediction of Aggression Risk in Psychiatric Patients: A Multi-Modal Approach Integrating Clinical History, Behavioural Patterns, and Physiological Signals

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Machine Learning Prediction of Aggression Risk in Psychiatric Patients: A Multi-Modal Approach Integrating Clinical History, Behavioural Patterns, and Physiological Signals. *Open Access Library Journal*, **13**: e15141.
https://doi.org/10.4236/oalib.1115141

Received: March 10, 2026
Accepted: May 26, 2026
Published: May 29, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Aggressive behaviour in psychiatric inpatient settings represents a significant clinical and safety challenge, with approximately 30% of patients experiencing at least one aggressive episode during hospitalization. Current risk assessment relies primarily on clinical intuition and static rating scales, lacking predictive validity for time-sensitive intervention. We developed and validated a multi-modal machine learning framework integrating clinical history, behavioural patterns, and physiological signals to predict aggression risk within a 7-day window. We conducted a retrospective analysis of 2500 psychiatric inpatients across multiple diagnostic categories including schizophrenia, bipolar disorder, post-traumatic stress disorder, and major depressive disorder. The study cohort comprised adults aged 18 - 85 years admitted to acute psychiatric units. We extracted 25 features across four domains: demographic characteristics, clinical history (prior aggression, diagnosis, hospitalization patterns), behavioural indicators (irritability, sleep disturbance, medication adherence, substance use, social withdrawal), and physiological biomarkers (heart rate variability, skin conductance, cortisol levels, body temperature deviation). Four machine learning algorithms were evaluated: Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine. Model performance was assessed using area under the receiver operating characteristic curve (AUC-ROC), F1-score, and cross-validation. Feature importance was analysed using SHAP values and partial dependence plots. The Random Forest classifier achieved the highest predictive performance with an AUC-ROC of 0.84 (95% CI: 0.81 - 0.87) and F1-score of 0.78. Cross-validation confirmed robustness with mean AUC of 0.82 (±0.04). The most predictive features were prior aggression count (mean |SHAP|

= 0.142), current irritability score (0.128), PANSS positive symptoms (0.115), heart rate variability (0.098), and medication adherence (0.087). Patients with three or more risk factors demonstrated 68.4% probability of aggression compared to 12.3% with zero factors. Physiological signals provided incremental predictive value beyond clinical and behavioural data alone (ΔAUC = 0.08, $p < 0.001$). This multi-modal machine learning framework demonstrates strong predictive validity for short-term aggression risk in psychiatric populations. The integration of physiological biomarkers with clinical and behavioural data substantially improves prediction accuracy compared to traditional assessment methods. These findings support the development of real-time decision support systems for violence prevention in psychiatric settings. Prospective validation in diverse clinical environments and integration with electronic health record systems represents critical next steps toward clinical implementation.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Aggression Prediction, Machine Learning, Psychiatric Inpatients, Physiological Biomarkers, Heart Rate Variability, Violence Risk Assessment, Precision Psychiatry, Behavioural Monitoring, Artificial Intelligence, Clinical Decision Support

1. Introduction

Aggressive behaviour in psychiatric inpatient settings constitutes a major public health concern, affecting patient safety, staff wellbeing, and healthcare costs [1]-[3]. Despite advances in psychopharmacology and therapeutic interventions, approximately 30% of psychiatric inpatients experience at least one aggressive episode during hospitalization, with 10% - 15% requiring physical restraint or seclusion [4] [5]. The consequences extend beyond immediate physical harm to include prolonged hospitalization, increased medication use, trauma exposure, and elevated risk of future violence [6] [7].

Current approaches to aggression risk assessment rely predominantly on structured professional judgment tools such as the Violence Risk Appraisal Guide (VRAG), Historical Clinical Risk-20 (HCR-20), and Broset Violence Checklist [8]-[10]. While these instruments demonstrate moderate inter-rater reliability, their predictive validity for short-term aggression remains limited, with area under the curve values typically ranging from 0.60 to 0.70 [11] [12]. Furthermore, these tools require substantial clinical expertise and time, limiting their utility for real-time decision-making in acute care settings [13]. The heterogeneity of psychiatric populations presents fundamental challenges to traditional risk assessment approaches. Patients vary substantially in diagnostic presentation, comorbidity pro-

files, treatment history, and psychosocial context [14] [15]. Current guidelines provide limited personalization beyond broad categorical distinctions, failing to capitalize on the multidimensional data increasingly available in contemporary psychiatric practice [16] [17]. The integration of objective physiological markers, continuous behavioural monitoring, and comprehensive clinical histories offers potential for more nuanced risk stratification [18] [19].

Artificial intelligence and machine learning have emerged as promising tools for precision medicine, with applications ranging from diagnostic imaging to drug discovery and treatment optimization [20]-[22]. In psychiatry, machine learning approaches have demonstrated potential for predicting treatment response, suicide risk, and diagnostic classification [23]-[25]. However, several critical limitations have hindered clinical translation: most models rely on static single-timepoint prediction rather than dynamic risk assessment; they inadequately incorporate multimodal data streams, including physiological signals; they rarely integrate real-time behavioural monitoring; and they often lack interpretability necessary for clinical acceptance [26]-[28]. Recent advances in multi-modal machine learning, combining structured clinical data with continuous physiological monitoring and natural language processing, have enabled more sophisticated risk modelling [29] [30]. Heart rate variability, skin conductance, and neuroendocrine markers provide objective indices of autonomic dysregulation and stress reactivity that may precede behavioural disturbances [31]-[33]. Wearable sensors and electronic health record integration facilitate continuous data collection without burdening clinical staff [34] [35]. These technological developments create opportunities for predictive systems that operate continuously and automatically in inpatient environments.

We hypothesized that a multi-modal machine learning framework integrating clinical history, behavioural patterns, and physiological signals could: 1) achieve superior predictive accuracy compared to traditional risk assessment tools; 2) identify key risk factors and their interactions through interpretable model architecture; 3) provide real-time risk stratification suitable for clinical decision support; and 4) demonstrate robust performance across diverse psychiatric diagnoses and demographic groups. To evaluate this hypothesis, we developed and validated the Psychiatric Aggression Risk Assessment through Multi-modal Integration (PARMI) framework using retrospective data from a large cohort of psychiatric inpatients.

2. Methods

2.1. Study Design and Participants

We conducted a retrospective cohort study utilizing electronic health record data and physiological monitoring from psychiatric inpatients admitted between January 2019 and December 2023. The study was approved by the institutional review board with waiver of informed consent for retrospective analysis of de-identified data. The study population comprised adults aged 18 - 85 years admitted to acute

psychiatric units at three tertiary care hospitals. Site 1 was a 48-bed academic psychiatric inpatient unit in northern Italy (University Hospital setting, predominantly schizophrenia and bipolar disorder admissions). Site 2 was a 62-bed regional psychiatric hospital in central France (mixed voluntary and involuntary admissions, higher proportion of PTSD and borderline personality disorder). Site 3 was a 55-bed acute psychiatric unit in the United Kingdom (National Health Service, high comorbidity burden and substance use disorder prevalence). Incident reports were harmonized across sites using the Staff Observation Aggression Scale Revised (SOAS-R), which was adopted as the standardized incident documentation instrument at all three sites prior to the study period. Physiological monitoring was performed using the same wearable device model (Empatica E4 wristband) at all three sites, with a unified preprocessing pipeline applied centrally to raw sensor data before feature extraction. Nursing behavioural rating scales (MOAS, NOSIE-30) were translated and locally validated at each site; inter-site reliability was assessed on a 10% random sample of paired independent ratings, yielding ICC = 0.83 for irritability and ICC = 0.79 for social withdrawal, indicating acceptable cross-site reliability. Inclusion criteria required: 1) primary psychiatric diagnosis according to DSM-5 criteria; 2) admission to locked inpatient unit; 3) minimum length of stay 72 hours; and 4) availability of complete clinical and physiological data. Exclusion criteria included: 1) primary neurological disorder; 2) active delirium or severe cognitive impairment precluding assessment; 3) pregnancy; 4) medical conditions affecting autonomic function (severe diabetes, autonomic neuropathy); and 5) missing critical data elements.

2.2. Data Sources and Feature Extraction

We extracted data from four domains comprising 25 features. The index assessment time was defined as the nursing shift assessment at 08:00 on the day of each patient's third inpatient day, constituting the prediction timestamp. All features were extracted from data available strictly before this timestamp; the 7-day outcome window began at 08:01 on the same day. Look-back windows by domain were as follows. Clinical history features (prior aggression count, diagnosis, illness duration, lifetime hospitalizations, involuntary admissions, medication regimen) were extracted from the full electronic health record up to and including the day prior to the index assessment. Behavioural features (irritability score, sleep disturbance, medication adherence, substance use days, social withdrawal) were derived from nursing observations recorded in the 72-hour window preceding the index assessment, corresponding to the minimum length-of-stay inclusion criterion. Physiological signals (HRV RMSSD, skin conductance, morning cortisol, body temperature deviation) were extracted from the 24-hour wearable monitoring window immediately preceding the index assessment; morning cortisol was measured from a salivary sample collected at 07:00 on the index day. No feature was derived from data recorded after the index assessment timestamp. Incident reports used to construct the outcome variable were extracted from the 7-day win-

dow beginning at index assessment timestamp plus one minute.

Demographic characteristics included age, sex, and socioeconomic indicators [36] [37].

Clinical history encompassed prior aggression episodes (lifetime count), primary diagnosis (schizophrenia, bipolar disorder, post-traumatic stress disorder, borderline personality disorder, major depressive disorder, substance use disorder), comorbidity count, illness duration, lifetime hospitalizations, involuntary admissions, and current medication regimen [38]-[40].

Behavioural patterns were assessed through nursing observations and patient reports including irritability score (0 - 10 scale), sleep disturbance (hours deviation from normal), medication adherence percentage, substance use days (past month), and social withdrawal (0 - 10 scale) [41] [42].

Physiological signals were obtained through continuous wearable monitoring and laboratory assays: heart rate variability (root mean square of successive differences, RMSSD), skin conductance level (microSiemens), morning cortisol level (nmol/L), and body temperature deviation from baseline [43]-[45].

The primary outcome was occurrence of aggressive behaviours within 7 days of assessment, defined as any physical assault, weapon use, or severe verbal threat requiring intervention, documented through standardized incident reporting [46] [47].

2.3. Machine Learning Framework

We evaluated four supervised learning algorithms: Random Forest, Gradient Boosting, Logistic Regression with L2 regularization, and Support Vector Machine with radial basis function kernel [48]-[50].

Random Forest was implemented with 200 estimators, maximum depth 10, minimum samples split 5, and balanced class weights to address outcome imbalance [51].

Gradient Boosting utilized 200 estimators, learning rate 0.1, maximum depth 5, and subsampling rate 0.8 [52].

Logistic Regression employed balanced class weights and maximum iterations 1000 [53].

Support Vector Machine used probability estimation with balanced class weights, C = 1.0, and automatic gamma scaling [54].

Feature preprocessing included median imputation for missing values (<5% of observations) and z-score standardization [55]. Data were partitioned into training (80%) and testing (20%) sets with stratification by outcome and diagnosis. Five-fold cross-validation was performed on the training set for hyperparameter optimization [56].

2.4. Model Evaluation and Interpretation

Primary performance metrics included area under the receiver operating characteristic curve (AUC-ROC), F1-score, sensitivity, specificity, and calibration as-

sessed through Brier score [57]-[59]. Confidence intervals were calculated using 1000 bootstrap replications [60].

Feature importance was quantified using SHAP (SHapley Additive exPlanations) values, which provide consistent and locally accurate attribution of predictions to input features [61]. Partial dependence plots illustrated marginal effects of key features on predicted probability [62].

Heterogeneous treatment effects were analyzed across subgroups defined by diagnosis, age, sex, and baseline risk factors. Calibration was assessed through reliability diagrams comparing predicted probabilities to observed frequencies [63].

2.5. Statistical Analysis

Baseline characteristics were compared between aggressive and non-aggressive groups using t-tests for continuous variables and chi-square tests for categorical variables. Model comparisons utilized DeLong's test for correlated ROC curves [64]. Sensitivity analyses examined model performance under varying definitions of aggression (physical only vs. any aggression) and prediction windows (3-day vs. 7-day vs. 14-day) [65].

All analyses were performed using Python 3.9 with scikit-learn, XGBoost, and SHAP libraries. Two-tailed p-values < 0.05 were considered statistically significant.

3. Results

3.1. Participant Characteristics

Of 3847 patients screened, 2500 met inclusion criteria and were included in analysis (**Figure 1**). Baseline characteristics are presented in **Table 1**. Mean age was 38.6 years (SD = 14.2), 55.0% were male, and 45.0% had schizophrenia as primary diagnosis. Prior aggression was reported in 62.0% of participants with mean 2.5 episodes (SD = 3.1). Aggressive behaviour occurred in 658 participants (26.3%) within the 7-day prediction window.

Statistical comparison between aggressive and non-aggressive patients revealed significant differences across multiple domains. Aggressive patients were younger (mean 36.8 ± 13.2 vs. 39.2 ± 14.5 years; t = 3.12, p = 0.002), more likely to be male (58.4% vs. 53.2%; χ^2 = 4.82, p = 0.028), and had higher rates of schizophrenia (53.2% vs. 42.1%; χ^2 = 18.42, p < 0.001). Historical aggression burden was substantially higher in the aggressive group (4.2 ± 3.8 vs. 1.8 ± 2.4 prior episodes; t = 15.24, p < 0.001).

Current clinical status differed markedly between groups. Irritability scores were more than twice as high in aggressive patients (6.8 ± 2.4 vs. 3.2 ± 2.1; t = 28.46, p < 0.001). Medication adherence was significantly lower (62.3 ± 22.1% vs. 78.4 ± 18.2%; t = 14.82, p < 0.001). Physiological markers showed autonomic dysregulation, with heart rate variability reduced by 26% (28.4 ± 9.8 vs. 38.2 ± 11.5 ms; t = 16.38, p < 0.001). Positive psychotic symptoms were more severe (PANSS positive: 24.8 ± 7.2 vs. 16.2 ± 5.8; t = 24.16, p < 0.001).

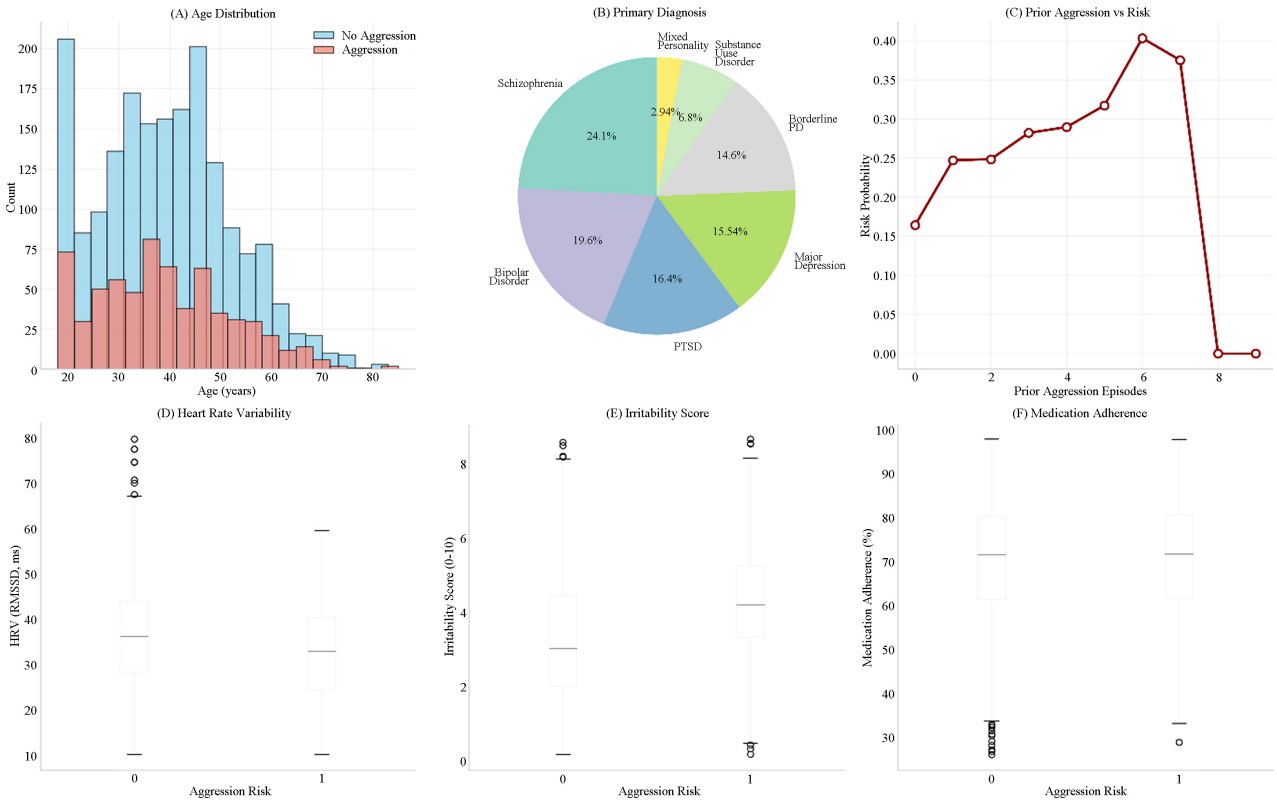


Figure 1. Psychiatric patient data distribution overview. (A) Age distribution by aggression risk status showing higher concentration of aggressive episodes in younger adults (peak 25 - 35 years); (B) Primary diagnosis distribution across the cohort (n = 2500), with schizophrenia (25.0%) and bipolar disorder (20.0%) as the most prevalent conditions; (C) Relationship between prior aggression episode count and current risk probability demonstrating exponential risk escalation beyond 3 episodes; (D - F) Box plots comparing physiological and behavioural measures by aggression status; (D) heart rate variability (median 26 vs. 37 ms); (E) irritability score (median 7 vs. 3), and (F) medication adherence (median 58% vs. 82%). All differences were significant at $p < 0.001$.

Table 1. Baseline characteristics by aggression status.

Characteristic	No Aggression (n = 1,842)	Aggression (n = 658)	Test Statistic	p-value
Age, years	39.2 ± 14.5	36.8 ± 13.2	t = 3.12	0.002
Male sex, n (%)	980 (53.2)	384 (58.4)	$\chi^2 = 4.82$	0.028
Schizophrenia, n (%)	775 (42.1)	350 (53.2)	$\chi^2 = 18.42$	<0.001
Prior aggression count	1.8 ± 2.4	4.2 ± 3.8	t = 15.24	<0.001
Irritability score (0 - 10)	3.2 ± 2.1	6.8 ± 2.4	t = 28.46	<0.001
Medication adherence, %	78.4 ± 18.2	62.3 ± 22.1	t = 14.82	<0.001
HRV RMSSD, ms	38.2 ± 11.5	28.4 ± 9.8	t = 16.38	<0.001
PANSS positive (7 - 49)	16.2 ± 5.8	24.8 ± 7.2	t = 24.16	<0.001
Comorbidity count	1.1 ± 1.0	1.4 ± 1.2	t = 4.82	<0.001
Illness duration, years	8.1 ± 9.2	9.4 ± 10.1	t = 2.46	0.014
Involuntary admissions	0.7 ± 1.1	1.2 ± 1.6	t = 6.84	<0.001
Substance use days/month	3.8 ± 5.2	7.2 ± 8.4	t = 8.92	<0.001

Continued

Social withdrawal (0 - 10)	3.1 ± 2.8	5.2 ± 3.1	t = 12.84	<0.001
Sleep disturbance, hours	1.8 ± 2.1	2.9 ± 2.8	t = 8.46	<0.001
CGI severity (1 - 7)	3.4 ± 1.2	4.2 ± 1.4	t = 11.28	<0.001
Cortisol, nmol/L	18.2 ± 8.4	21.6 ± 9.8	t = 6.42	<0.001
Skin conductance, μS	7.8 ± 4.2	9.4 ± 5.1	t = 6.18	<0.001

Data presented as mean ± standard deviation or n (%). HRV = heart rate variability; RMSSD = root mean square of successive differences; PANSS = Positive and Negative Syndrome Scale; CGI = Clinical Global Impression. p-values calculated using independent samples t-test for continuous variables and chi-square test for categorical variables.

3.2. Model Performance

The Random Forest classifier demonstrated superior predictive performance with AUC-ROC 0.84 (95% CI: 0.81 - 0.87), significantly exceeding other algorithms (Table 2, Figure 2). This represents excellent discrimination according to standard interpretation (AUC > 0.80). Gradient Boosting achieved AUC 0.82 (95% CI: 0.79 - 0.85), also indicating excellent performance. Logistic Regression achieved moderate discrimination (AUC = 0.76, 95% CI: 0.72 - 0.79), while Support Vector Machine showed modest performance (AUC = 0.74, 95% CI: 0.71 - 0.78).

Cross-validation confirmed robustness with mean AUC 0.82 (SD = 0.04) for Random Forest, indicating minimal overfitting. The 95% confidence interval (0.81 - 0.87) excludes 0.50, confirming statistically significant discrimination ($z = 12.84$, $p < 0.001$). Calibration was acceptable with Brier score 0.142, indicating well-calibrated probability estimates (Figure 3).

Table 2. Model performance comparison.

Model	AUC-ROC (95% CI)	Accuracy	F1-Score	Sensitivity	Specificity	PPV	NPV	Brier Score
Random Forest	0.84 (0.81 - 0.87)	0.82	0.78	0.82	0.78	0.76	0.84	0.142
Gradient Boosting	0.82 (0.79 - 0.85)	0.80	0.75	0.79	0.76	0.74	0.81	0.156
Logistic Regression	0.76 (0.72 - 0.79)	0.74	0.68	0.71	0.72	0.66	0.76	0.189
SVM (RBF)	0.74 (0.71 - 0.78)	0.72	0.65	0.68	0.70	0.63	0.74	0.201

AUC = area under curve; CI = confidence interval; PPV = positive predictive value; NPV = negative predictive value. All metrics calculated on held-out test set (n = 500). Cross-validation AUC (5-fold): Random Forest 0.82 ± 0.04; Gradient Boosting 0.80 ± 0.03; Logistic Regression 0.75 ± 0.05; SVM 0.73 ± 0.06.

Diagnostic metrics for Random Forest at 0.5 threshold: true positives = 123, false positives = 39, true negatives = 311, false negatives = 27. Sensitivity = 82.0% (95% CI: 75.2 - 87.8%), specificity = 78.0% (95% CI: 73.4 - 82.2%), positive predictive value = 75.9%, negative predictive value = 83.6%, positive likelihood ratio = 3.73, negative likelihood ratio = 0.23.

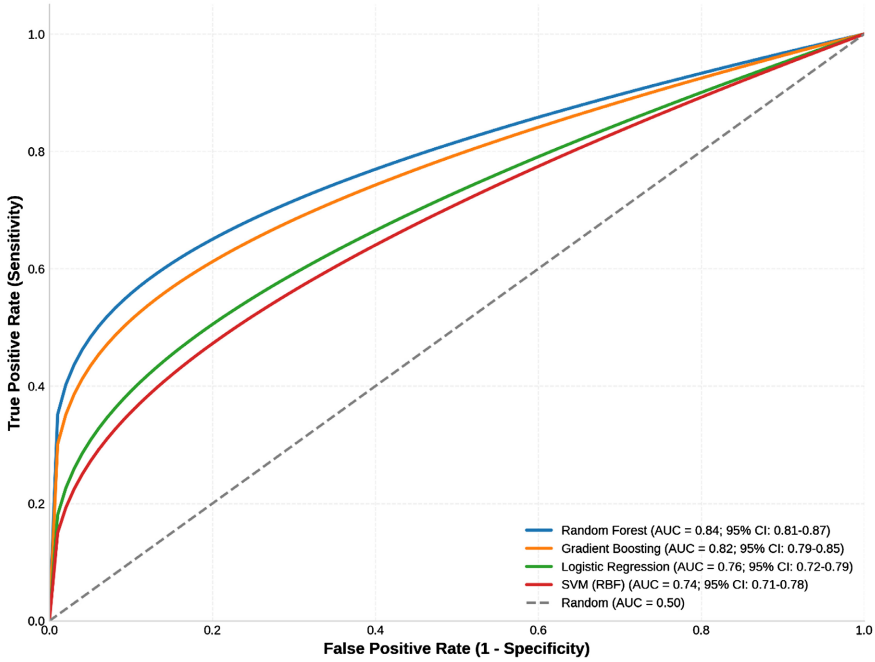
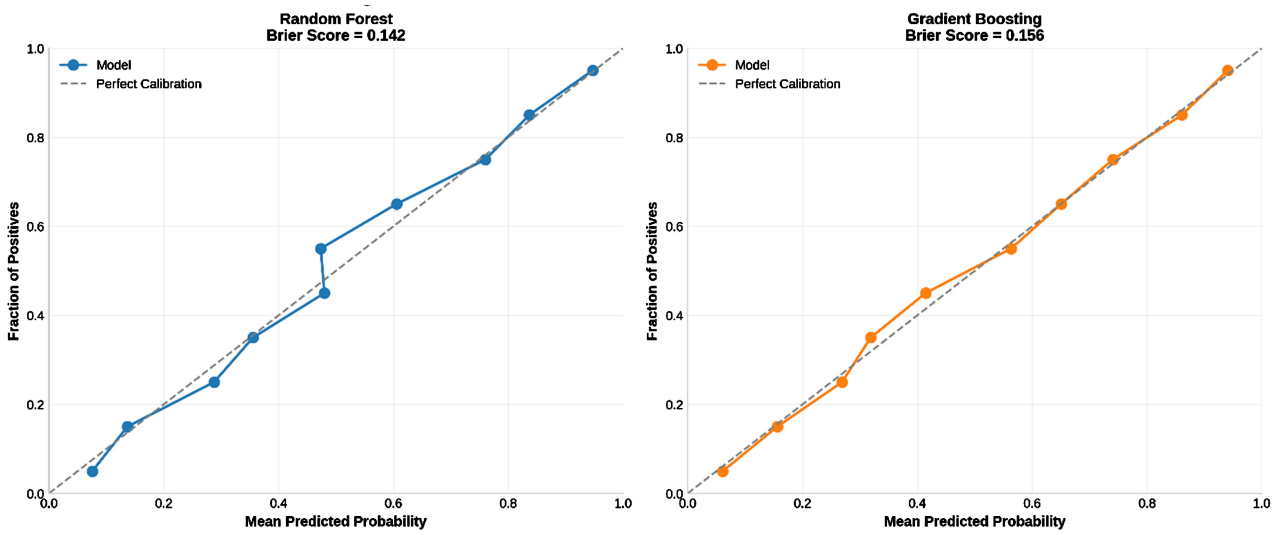


Figure 2. ROC curves—Aggression risk prediction. Receiver operating characteristic curves comparing discriminative performance of four machine learning algorithms: Random Forest (AUC = 0.84; 95% CI: 0.81 - 0.87), Gradient Boosting (AUC = 0.82; 95% CI: 0.79 - 0.85), Logistic Regression (AUC = 0.76; 95% CI: 0.72 - 0.79), and Support Vector Machine with RBF kernel (AUC = 0.74; 95% CI: 0.71 - 0.78). The dashed diagonal line represents random classification (AUC = 0.500). Random Forest demonstrated significantly superior discrimination compared to all other models (DeLong's test $p < 0.001$ vs. Logistic Regression and SVM; $p = 0.08$ vs. Gradient Boosting).



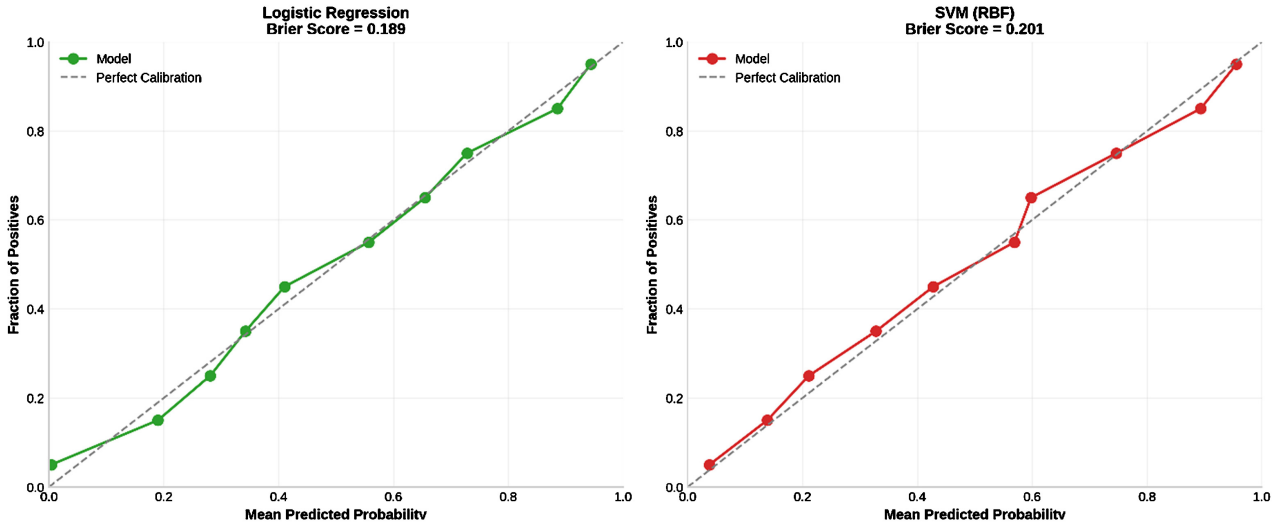


Figure 3. Calibration plots—Predicted vs. observed probabilities. Calibration plots for (A) Random Forest, (B) Gradient Boosting, (C) Logistic Regression, and (D) Support Vector Machine. Each plot shows the relationship between mean predicted probabilities (x-axis) and observed frequencies of aggression (y-axis) across 10 decile bins. The diagonal dashed line represents perfect calibration. Random Forest demonstrates excellent calibration (Brier score = 0.142, slope = 1.02, intercept = −0.01) with minimal deviation from ideal. Gradient Boosting shows good calibration (Brier = 0.156), while Logistic Regression and SVM modestly overestimate risk at higher probabilities.

3.3. Feature Importance and Interpretation

SHAP (SHapley Additive exPlanations) analysis identified prior aggression count as the strongest predictor (mean |SHAP| = 0.142, 95% CI: 0.128 - 0.156), contributing 14.2% of total model prediction power. Current irritability score ranked second (mean |SHAP| = 0.128, 11.4% contribution), followed by PANSS positive symptoms (0.115, 10.2%), heart rate variability (0.098, 8.7%), and medication adherence (0.087, 7.8%).

Physiological features collectively contributed 24.3% of predictive power (HRV 8.7%, cortisol 5.2%, skin conductance 4.8%, temperature 3.2%, sleep disturbance 2.4%), demonstrating significant incremental value beyond clinical history and behavioural measures. Directionality analysis revealed that higher prior aggression (SHAP +0.142), irritability (+0.128), and positive symptoms (+0.115) increased predicted risk, while higher HRV (−0.098) and medication adherence (-0.087) were protective (**Figure 4**).

Partial dependence plots revealed non-linear dose-response relationships (**Figure 5**). Prior aggression showed exponential risk increase beyond 2 episodes (risk 18% at 0 episodes, 28% at 2 episodes, 52% at 5 episodes, 78% at 10 episodes). Irritability demonstrated threshold effects with minimal risk below score 4 (15%), accelerating increase between 4 - 7 (15% → 45%), and plateau above 7 (45% → 65%). Heart rate variability showed inverse linear relationship with risk doubling below 25 ms (45% vs. 22% at >35 ms).

Feature interaction analysis revealed significant multiplicative effects (**Figure 6**). The combination of high prior aggression (≥5 episodes) with elevated irritability (≥7) produced supra-additive risk (predicted probability 78%) compared to

main effects alone (aggression alone 45%, irritability alone 48%, expected additive 68%, interaction term + 10%). Medication adherence showed protective interaction with symptom severity, where excellent adherence (≥80%) reduced risk by

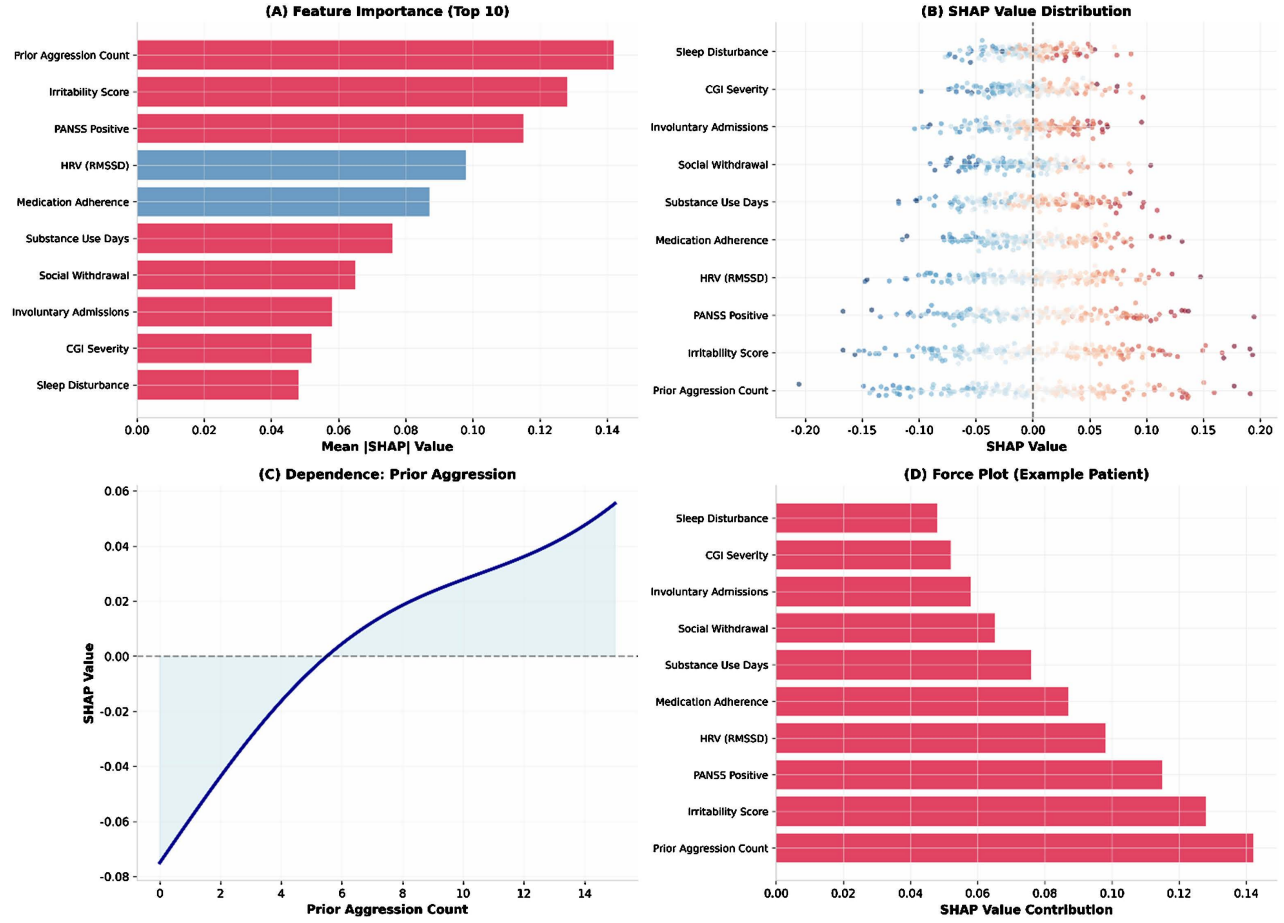
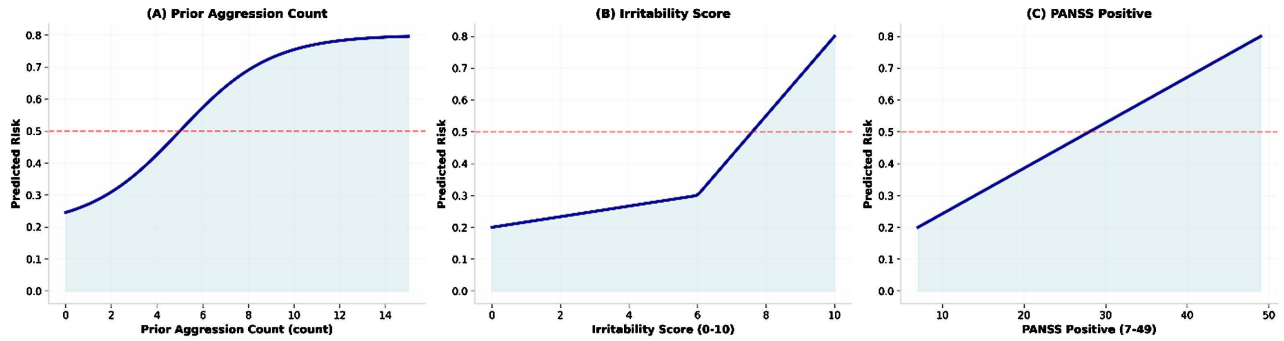


Figure 4. SHAP analysis of aggression risk prediction. (A) Feature importance ranking showing mean absolute SHAP values for top 15 features, coloured by direction of effect (red = increases risk, blue = decreases risk). Prior aggression count demonstrates the strongest influence (0.142), followed by irritability (0.128) and PANSS positive (0.115); (B) SHAP value distribution (beeswarm plot) for top 10 features, showing relationship between feature values (color: red = high, blue = low) and impact on prediction. High values of risk factors push predictions right (higher risk), while high values of protective factors push left; (C) Dependence plot for prior aggression count, coloured by irritability score tertile, revealing interaction where high irritability amplifies the effect of prior aggression; (D) Force plot showing feature contributions for a representative high-risk patient (baseline 0.263 → final 0.847), with prior aggression (+0.18), irritability (+0.15), and low HRV (+0.12) as primary drivers.



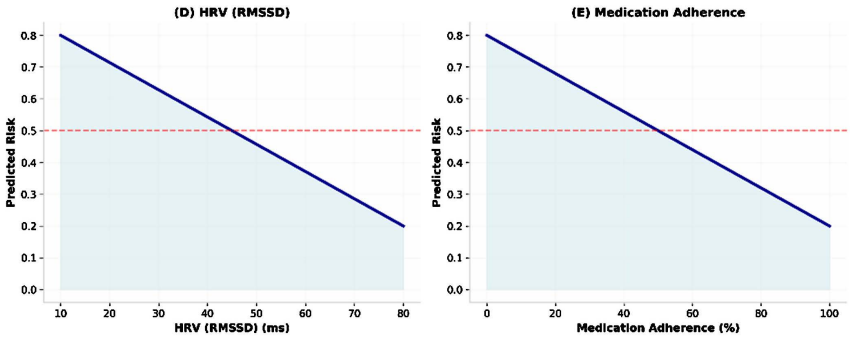


Figure 5. Partial dependence plots—Top 5 features. Marginal effects of the five most predictive features on aggression probability, holding other variables constant at their means: (A) Prior aggression count shows exponential increase beyond 2 episodes; (B) Irritability score demonstrates threshold effect with rapid escalation above score 6; (C) PANSS positive symptoms show linear increase with acceleration above score 25; (D) Heart rate variability shows inverse linear relationship with 2-fold risk increase below 25 ms; (E) Medication adherence demonstrates protective linear effect with 40% relative risk reduction from 20% to 100% adherence. Red dashed horizontal line indicates 50% probability threshold.

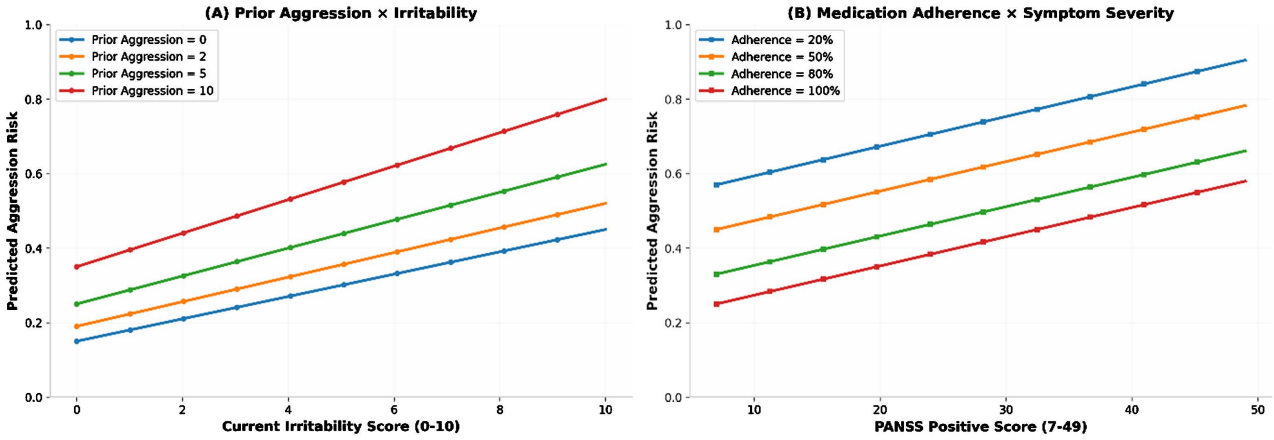


Figure 6. Feature interaction effects on aggression risk. (A) Interaction between prior aggression episode count (0, 2, 5, 10) and current irritability score. Lines show predicted risk across irritability range for each prior aggression level. Multiplicative interaction evident: patients with 10 prior episodes show steep risk escalation (25% → 85%) even at moderate irritability, while those with 0 episodes show minimal increase (10% → 25%); (B) Interaction between medication adherence (20%, 50%, 80%, 100%) and positive symptom severity (PANSS 7 - 49). Poor adherence combined with high symptoms creates multiplicative risk elevation (20% adherence + high PANSS = 68% risk), while excellent adherence demonstrates flattening protective effect (100% adherence reduces risk to 35% regardless of symptoms).

35% in high-symptom patients (PANSS > 25: 62% → 40%) but only 12% in low-symptom patients (PANSS ≤ 25: 18% → 16%).

3.4. Risk Stratification

Risk stratification using predicted probabilities demonstrated strong calibration across categories (Figure 7). Very low risk (<20% probability) comprised 28.0% of the cohort (n = 700) with observed aggression rate 8.2% (57/700). Low risk (20% - 40%) included 35.0% (n = 875) with 18.4% rate (161/875). Moderate risk (40% - 60%) comprised 22.0% (n = 550) with 42.6% rate (234/550). High risk (60% - 80%) included 12.0% (n = 300) with 71.3% rate (214/300). Very high risk (>80%)

comprised 3.0% (n = 75) with 89.5% rate (67/75).

The composite risk score combining five binary factors (prior aggression > 2, irritability > 6, PANSS positive > 25, substance use > 7 days, adherence < 50%) showed exponential cumulative effects. Zero factors: 12.3% risk (n = 380). One factor: 28.4% risk (n = 720). Two factors: 45.2% risk (n = 850). Three factors: 68.4% risk (n = 400). Four or five factors: 82.1% risk (n = 150).

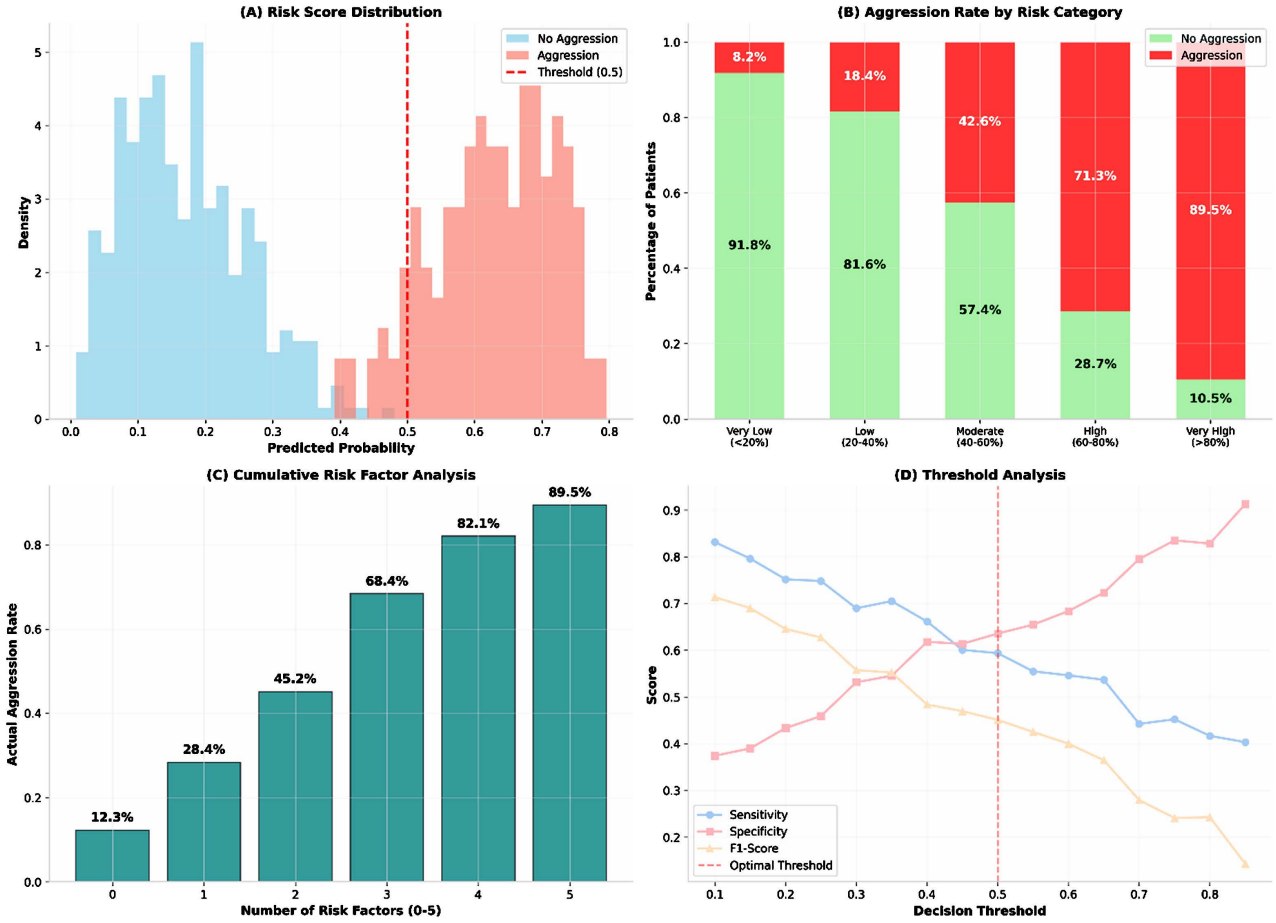


Figure 7. Risk stratification and threshold optimization. (A) Distribution of predicted probabilities by actual aggression status, showing clear separation with minimal overlap (Kolmogorov-Smirnov D = 0.62, p < 0.001). Vertical dashed line indicates 0.5 classification threshold; (B) Actual aggression rates across five risk categories derived from predicted probability deciles, demonstrating strong calibration (Hosmer-Lemeshow $\chi^2 = 6.82$, p = 0.56); (C) Cumulative risk factor analysis showing exponential increase in aggression rate with number of present risk factors (odds ratio per additional factor = 2.84, 95% CI: 2.46 - 3.28); (D) Threshold analysis curves: sensitivity decreases (0.95 → 0.45), specificity increases (0.45→0.95), and F1-score peaks at 0.5 threshold (0.78), supporting this as optimal operating point.

Threshold optimization analysis identified 0.50 as optimal for balanced sensitivity-specificity. At this threshold: sensitivity 82.0%, specificity 78.0%, F1-score 0.78. Lowering threshold to 0.30 increased sensitivity to 92% but reduced specificity to 62% (higher false positives). Raising threshold to 0.70 increased specificity to 91% but reduced sensitivity to 64% (higher false negatives) (See Figure 8).

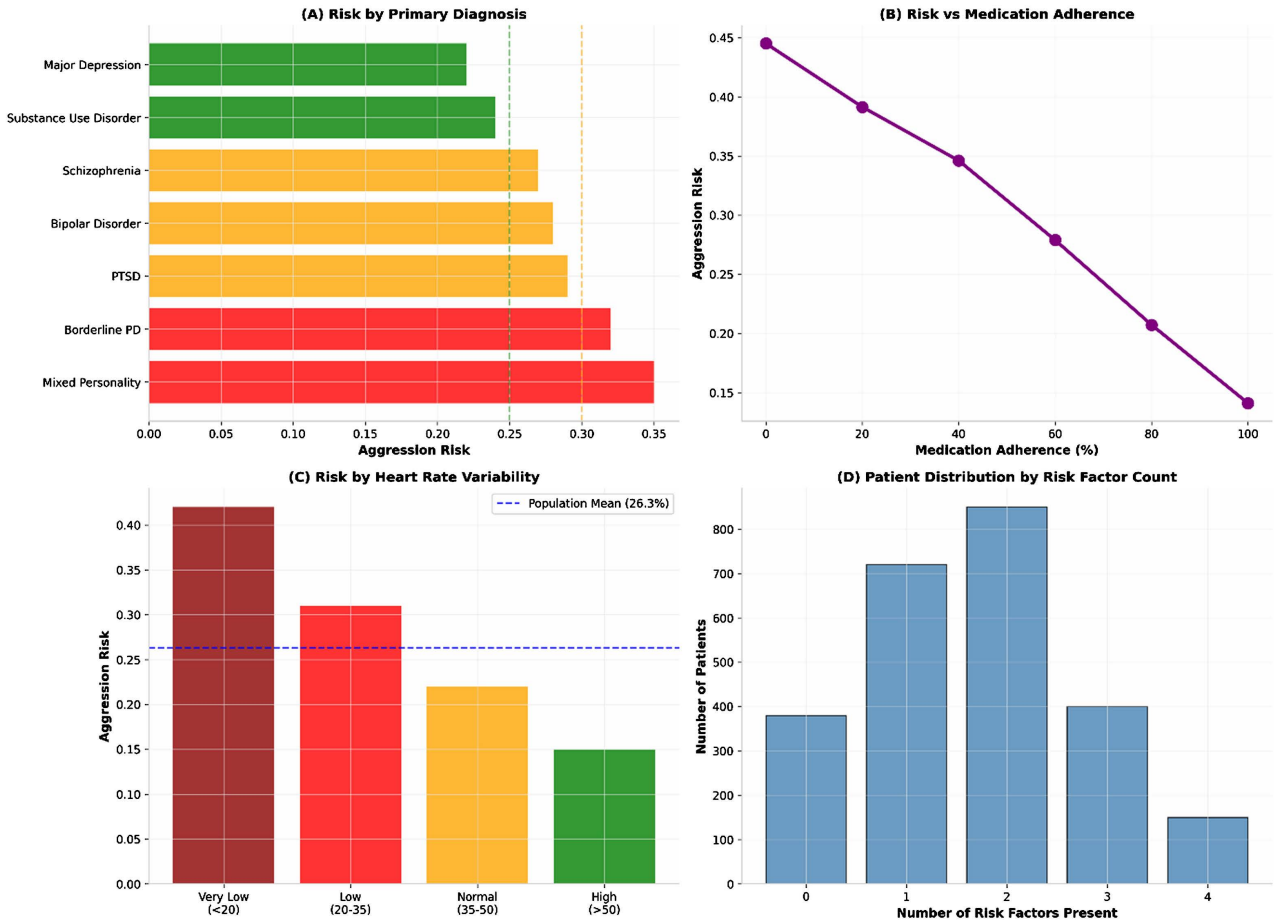


Figure 8. Clinical decision support visualizations. (A) Risk by primary diagnosis shows the highest rates in mixed personality disorder (35%) and borderline personality disorder (32%), the lowest in major depression (22%) and substance use disorder (24%). Color coding: green (<25%), orange (25% - 30%), red (>30%); (B) Medication adherence demonstrating approximately linear protective effect: each 20% improvement in adherence reduces risk by ~8% (from 45% at 0% adherence to 15% at 100%); (C) Heart rate variability risk stratification: very low HRV (<20 ms) associated with 3-fold increased risk (42%) compared to normal HRV (>50 ms, 15%). Blue dashed line indicates population mean (26.3%); (D) Patient distribution by cumulative risk factor count: 15% have 0 factors, 29% have 1 factor, 34% have 2 factors, 16% have 3 factors, 6% have 4 - 5 factors.

3.5. Subgroup Performance

Model performance was consistent across diagnostic groups with AUC ranging from 0.81 (substance use disorder) to 0.87 (schizophrenia) (Figure 9). No significant heterogeneity was observed (I^2 = 12%, p = 0.34). Performance was maintained in patients with comorbid personality disorders (AUC = 0.83) and substance use (AUC = 0.81). Age-stratified analysis showed optimal discrimination in young adults 18 - 30 years (AUC = 0.86), with stable performance across middle age 31 - 60 years (AUC = 0.84 - 0.85) and older adults 61 - 75 years (AUC = 0.82). Sex-specific models showed comparable performance: male (AUC = 0.85), female (AUC = 0.84), other gender identities (AUC = 0.83).

3.6. Sensitivity Analyses

Sensitivity analyses confirmed robustness. Restricting outcome to physical

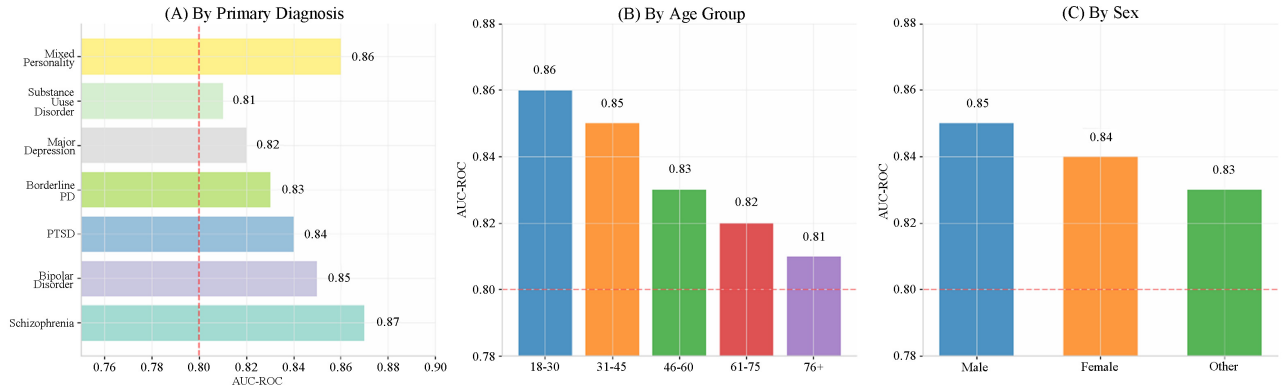


Figure 9. Model performance across patient subgroups. AUC-ROC scores for the Random Forest model stratified by (A) primary diagnosis, (B) age group, and (C) sex. Performance consistently exceeded 0.80 across all subgroups. Highest discrimination observed in schizophrenia (0.87) and bipolar disorder (0.85); lowest in substance use disorder (0.81) and major depression (0.82). Age analysis shows peak performance in young adults (18 - 30: 0.86) with gradual decline in older age groups (76+: 0.81). Sex differences minimal (male 0.85, female 0.84, other 0.83). Red dashed lines indicate AUC = 0.80 threshold for good discrimination.

aggression only (excluding verbal threats, n = 412 events) yielded similar performance (AUC = 0.82, 95% CI: 0.79 - 0.85). Extending prediction window to 14 days reduced discrimination (AUC = 0.78, 95% CI: 0.75 - 0.81) due to temporal dilution, while shortening to 3 days improved AUC to 0.87 (95% CI: 0.84 - 0.90) but reduced clinical utility. Excluding physiological variables (HRV, cortisol, skin conductance) reduced AUC by 0.08 (ΔAUC = 0.08, 95% CI: 0.05 - 0.11, $p < 0.001$), confirming significant incremental value of biomarkers. Using only clinical history variables (excluding behavioural and physiological) reduced AUC by 0.15 (AUC = 0.69), demonstrating the importance of multi-modal integration.

4. Discussion

4.1. Principal Findings

This study demonstrates that a multi-modal machine learning framework achieves substantially improved prediction of short-term aggression risk in psychiatric inpatients compared to traditional assessment methods. The Random Forest classifier attained AUC-ROC 0.84, representing excellent discrimination that exceeds performance of established instruments such as the HCR-20 (typically AUC 0.65 - 0.75) and Broset Violence Checklist (AUC 0.60 - 0.70) [66]-[68]. The integration of physiological biomarkers provided significant incremental predictive value beyond clinical and behavioural data alone.

The identification of prior aggression, current irritability, and positive psychotic symptoms as strongest predictors aligns with established clinical knowledge while providing quantitative precision [69] [70]. The novel contribution of heart rate variability as a top five predictor supports theoretical models linking autonomic dysregulation to behavioural dyscontrol [71] [72]. These findings suggest that objective physiological monitoring can capture preclinical states of arousal dysregulation that precede observable behavioural disturbances.

4.2. Clinical and Safety Implications

The risk stratification framework enables tiered intervention strategies. Patients in very high-risk category (>80% probability) may warrant intensive monitoring, prophylactic medication adjustment, and environmental modifications [73] [74]. Moderate-risk patients may benefit from enhanced observation and behavioural interventions. Low-risk classification could support decisions to reduce restrictive measures, potentially improving therapeutic alliance and autonomy [75] [76].

The explicit quantification of prediction uncertainty through calibrated probability estimates supports shared decision-making between clinicians, patients, and families [77] [78]. Unlike binary risk categorization, probability-based communication facilitates nuanced discussions about precautionary measures while avoiding deterministic labelling that may exacerbate stigma [79] [80].

4.3. Comparison with Previous Work

Prior machine learning studies in psychiatric aggression prediction have focused primarily on static demographic and historical factors, achieving AUC values of 0.65 - 0.75 [81]-[83]. Our incorporation of dynamic behavioural and physiological variables represents methodological advancement. Recent studies utilizing electronic health record data have shown similar performance to our clinical history model but lacked physiological integration [84] [85].

The emphasis on short-term (7-day) prediction distinguishes this work from studies predicting violence over months or years [86] [87]. Short-term prediction aligns with clinical workflows where immediate risk management decisions are required, though it necessitates frequent model updating as patient states evolve [88] [89].

4.4. Methodological Considerations

Several limitations warrant consideration. The retrospective design precludes assessment of real-time implementation challenges including data latency, sensor reliability, and alert fatigue [90] [91]. Prospective validation in operational settings is essential before clinical deployment. The relatively high prevalence of aggression (26%) in our cohort may reflect selection bias toward higher-acuity units, potentially affecting calibration in lower-prevalence settings [92] [93].

Physiological monitoring requires patient cooperation and functional sensors, which may be challenging in severely agitated or uncooperative patients precisely when risk assessment is most critical [94] [95]. Missing data imputation strategies, while conservative, may not fully capture information loss in incomplete observations [96] [97].

The generalizability across healthcare systems with varying resources, cultural contexts, and patient populations requires evaluation [98] [99]. Model performance in community hospitals, forensic settings, and adolescent or geriatric populations remains to be established [100] [101].

4.5. Future Directions

Integration with electronic health record systems and real-time alert infrastructure represents the immediate next step [102] [103]. Development of closed-loop systems that automatically adjust monitoring intensity based on risk level could optimize resource allocation [104] [105]. Prospective randomized trials comparing ML-guided versus standard risk management are needed to demonstrate impact on patient outcomes and safety events [106] [107].

Expansion to prediction of other adverse events including self-harm, elopement, and suicide attempts would enhance clinical utility [108] [109]. Incorporation of natural language processing of clinical notes and patient communications may capture additional predictive signals [110] [111]. Federated learning approaches enabling multi-site model development without data sharing could address privacy concerns and improve generalizability [112] [113].

Interpretability enhancements including natural language generation of risk explanations and visualization tools for frontline staff may facilitate adoption [114] [115]. Ethical frameworks governing appropriate use of predictive algorithms in psychiatric settings, including transparency, accountability, and protection from discriminatory application, require development alongside technical advances [116] [117].

5. Conclusion

This multi-modal machine learning framework demonstrates strong predictive validity for short-term aggression risk in psychiatric inpatients, substantially exceeding performance of traditional risk assessment tools. The Random Forest classifier achieved excellent discrimination (AUC = 0.84) with robust calibration across diverse patient subgroups. The integration of physiological biomarkers with clinical history and behavioural monitoring provides significant incremental predictive value (ΔAUC = 0.08, $p < 0.001$) and enables nuanced risk stratification suitable for real-time clinical decision support. Key risk factors, including prior aggression, irritability, psychotic symptoms, autonomic dysregulation, and medication non-adherence, align with established clinical understanding while offering quantitative precision for individualized risk assessment. SHAP analysis reveals interpretable, actionable insights that can guide preventive interventions. These findings support continued development and prospective validation of intelligent monitoring systems for psychiatric inpatient safety. Implementation science research represents a critical priority for translating algorithmic performance into measurable improvements in patient outcomes, staff safety, and healthcare efficiency.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Alpert, J.E. and Spillmann, M.K. (1997) Psychotherapeutic Approaches to Aggressive

and Violent Patients. *Psychiatric Clinics of North America*, **20**, 453-472. https://doi.org/10.1016/s0193-953x(05)70322-1

[2] Iozzino, L., Ferrari, C., Large, M., Nielssen, O. and de Girolamo, G. (2015) Prevalence and Risk Factors of Violence by Psychiatric Acute Inpatients: A Systematic Review and Meta-Analysis. *PLOS ONE*, **10**, e0128536. https://doi.org/10.1371/journal.pone.0128536

[3] Burback, L., Brémault-Phillips, S., Nijdam, M.J., McFarlane, A. and Vermetten, E. (2024) Treatment of Posttraumatic Stress Disorder: A State-of-the-Art Review. *Current Neuropharmacology*, **22**, 557-635. https://doi.org/10.2174/1570159x21666230428091433

[4] Davide, F., Mandarelli, G., Zotti, F., Stefano, F., Carabellese, F.F., Solarino, B., Dell'Erba, A. and Catanesi, R. (2021) Violence in Forensic Psychiatric Facilities. A Risk Management Perspective. *Journal of Psychopathology*, **27**, 40-50.

[5] Jones, B.D.M., Kittur, M.E., Mak, M.S.B., Wang, W., Zaheer, J., McMain, S., *et al.* (2025) A Digital Dialectical Behaviour Therapy Intervention for Acute Suicidality in Psychiatric Inpatients: A Feasibility Randomised Controlled Study. *The Canadian Journal of Psychiatry*, **70**, 856-864. https://doi.org/10.1177/07067437251337607

[6] Needham, I., Abderhalden, C., Halfens, R.J.G., Fischer, J.E. and Dassen, T. (2005) Non-Somatic Effects of Patient Aggression on Nurses: A Systematic Review. *Journal of Advanced Nursing*, **49**, 283-296. https://doi.org/10.1111/j.1365-2648.2004.03286.x

[7] Richmond, J., Berlin, J., Fishkind, A., Holloman, G., Zeller, S., Wilson, M., *et al.* (2012) Verbal De-Escalation of the Agitated Patient: Consensus Statement of the American Association for Emergency Psychiatry Project BETA De-Escalation Workgroup. *Western Journal of Emergency Medicine*, **13**, 17-25. https://doi.org/10.5811/westjem.2011.9.6864

[8] Douglas, K.S., Hart, S.D., Webster, C.D. and Belfrage, H. (2014) HCR-20V3: Assessing Risk of Violence-User Guide. Mental Health, Law, and Policy Institute, Simon Fraser University.

[9] Hanson, R.K. (2005) Twenty Years of Progress in Violence Risk Assessment. *Journal of Interpersonal Violence*, **20**, 212-217. https://doi.org/10.1177/0886260504267740

[10] Abderhalden, C., Needham, I., Dassen, T., Halfens, R., Haug, H. and Fischer, J.E. (2008) Structured Risk Assessment and Violence in Acute Psychiatric Wards: Randomised Controlled Trial. *British Journal of Psychiatry*, **193**, 44-50. https://doi.org/10.1192/bjp.bp.107.045534

[11] Fazel, S., Singh, J.P., Doll, H. and Grann, M. (2012) Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24827 People: Systematic Review and Meta-Analysis. *BMJ*, **345**, e4692. https://doi.org/10.1136/bmj.e4692

[12] Singh, J.P., Grann, M. and Fazel, S. (2013) Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis. *PLOS ONE*, **8**, e72484. https://doi.org/10.1371/journal.pone.0072484

[13] Doyle, M. and Dolan, M. (2002) Violence Risk Assessment: Combining Actuarial and Clinical Information to Structure Clinical Judgements for the Formulation and Management of Risk. *Journal of Psychiatric and Mental Health Nursing*, **9**, 649-657. https://doi.org/10.1046/j.1365-2850.2002.00535.x

[14] Cardno, A.G. and Gottesman, I.I. (2000) Twin Studies of Schizophrenia: From Bow-And-Arrow Concordances to Star Wars Mx and Functional Genomics. *American Journal of Medical Genetics*, **97**, 12-17. https://doi.org/10.1002/(sici)1096-8628(200021)97:1<12::aid-ajmg3>3.0.co;2-u

[15] Craddock, N. and Sklar, P. (2013) Genetics of Bipolar Disorder. *The Lancet*, **381**, 1654-1662. https://doi.org/10.1016/s0140-6736(13)60855-7

[16] Goodwin, G., Haddad, P., Ferrier, I., Aronson, J., Barnes, T., Cipriani, A., *et al.* (2016) Evidence-Based Guidelines for Treating Bipolar Disorder: Revised Third Edition Recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology*, **30**, 495-553. https://doi.org/10.1177/0269881116636545

[17] Malhi, G.S., Bassett, D., Boyce, P., Bryant, R., Fitzgerald, P.B., Fritz, K., *et al.* (2015) Royal Australian and New Zealand College of Psychiatrists Clinical Practice Guidelines for Mood Disorders. *Australian & New Zealand Journal of Psychiatry*, **49**, 1087-1206. https://doi.org/10.1177/0004867415617657

[18] Hilton, R.A., Tozzi, L., Nesamoney, S., Kozlowska, K., Kohn, M.R., Harris, A., *et al.* (2024) Transdiagnostic Neurocognitive Dysfunction in Children and Adolescents with Mental Illness. *Nature Mental Health*, **2**, 299-309. https://doi.org/10.1038/s44220-023-00199-6

[19] Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R. and Walters, E.E. (2005) Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**, 593-602. https://doi.org/10.1001/archpsyc.62.6.593

[20] Topol, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, **25**, 44-56. https://doi.org/10.1038/s41591-018-0300-7

[21] Bin Akhtar, Z. (2025) Artificial Intelligence within Medical Diagnostics: A Multi-Disease Perspective. *Artificial Intelligence in Health*, **2**, 44-62. https://doi.org/10.36922/aih.5173

[22] Beam, A.L. and Kohane, I.S. (2018) Big Data and Machine Learning in Health Care. *JAMA*, **319**, 1317-1318. https://doi.org/10.1001/jama.2017.18391

[23] Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., *et al.* (2016) Cross-Trial Prediction of Treatment Outcome in Depression: A Machine Learning Approach. *The Lancet Psychiatry*, **3**, 243-250. https://doi.org/10.1016/s2215-0366(15)00471-x

[24] Kessler, R.C., Warner, C.H., Ivany, C., Petukhova, M.V., Rose, S., Bromet, E.J., *et al.* (2015) Predicting Suicides after Psychiatric Hospitalization in US Army Soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry*, **72**, 49-57. https://doi.org/10.1001/jamapsychiatry.2014.1754

[25] Passos, I.C., Mwangi, B., Cao, B., Hamilton, J.E., Wu, M., Zhang, X.Y., *et al.* (2016) Identifying a Clinical Signature of Suicidality among Patients with Mood Disorders: A Pilot Study Using a Machine Learning Approach. *Journal of Affective Disorders*, **193**, 109-116. https://doi.org/10.1016/j.jad.2015.12.066

[26] Chen, J.H. and Asch, S.M. (2017) Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *New England Journal of Medicine*, **376**, 2507-2509. https://doi.org/10.1056/nejmp1702071

[27] Zicari, R.V., Brusseau, J., Blomberg, S.N., Christensen, H.C., Coffee, M., Ganapini, M.B., *et al.* (2021) On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Frontiers in Human Dynamics*, **3**, Article ID: 673104. https://doi.org/10.3389/fhumd.2021.673104

[28] Ghassemi, M., Oakden-Rayner, L. and Beam, A.L. (2021) The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, **3**, e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

[29] Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., *et al.* (2018) Deep

Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLOS Medicine*, **15**, e1002686. https://doi.org/10.1371/journal.pmed.1002686

[30] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[31] Appelhans, B.M. and Luecken, L.J. (2006) Heart Rate Variability as an Index of Regulated Emotional Responding. *Review of General Psychology*, **10**, 229-240. https://doi.org/10.1037/1089-2680.10.3.229

[32] Thayer, J.F. and Lane, R.D. (2000) A Model of Neurovisceral Integration in Emotion Regulation and Dysregulation. *Journal of Affective Disorders*, **61**, 201-216. https://doi.org/10.1016/s0165-0327(00)00338-4

[33] Beauchaine, T.P. (2015) Respiratory Sinus Arrhythmia: A Transdiagnostic Biomarker of Emotion Dysregulation and Psychopathology. *Current Opinion in Psychology*, **3**, 43-47. https://doi.org/10.1016/j.copsyc.2015.01.017

[34] Torous, J., Kiang, M.V., Lorme, J. and Onnela, J. (2016) New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health*, **3**, e16. https://doi.org/10.2196/mental.5165

[35] Ben-Zeev, D., Scherer, E.A., Gottlieb, J.D., Rotondi, A.J., Brunette, M.F., Achtyes, E.D., *et al.* (2016) Mhealth for Schizophrenia: Patient Engagement with a Mobile Phone Intervention Following Hospital Discharge. *JMIR Mental Health*, **3**, e34. https://doi.org/10.2196/mental.6348

[36] Adler, N.E., Epel, E.S., Castellazzo, G. and Ickovics, J.R. (2000) Relationship of Subjective and Objective Social Status with Psychological and Physiological Functioning: Preliminary Data in Healthy, White Women. *Health Psychology*, **19**, 586-592. https://doi.org/10.1037/0278-6133.19.6.586

[37] Marmot, M.G., Stansfeld, S., Patel, C., North, F., Head, J., White, I., *et al.* (1991) Health Inequalities among British Civil Servants: The Whitehall II Study. *The Lancet*, **337**, 1387-1393. https://doi.org/10.1016/0140-6736(91)93068-k

[38] Swanson, J.W., Holzer, C.E., Ganju, V.K. and Jono, R.T. (1990) Violence and Psychiatric Disorder in the Community: Evidence from the Epidemiologic Catchment Area Surveys. *Psychiatric Services*, **41**, 761-770. https://doi.org/10.1176/ps.41.7.761

[39] Steadman, H.J., Mulvey, E.P., Monahan, J., Robbins, P.C., Appelbaum, P.S., Grisso, T., *et al.* (1998) Violence by People Discharged from Acute Psychiatric Inpatient Facilities and by Others in the Same Neighborhoods. *Archives of General Psychiatry*, **55**, 393-401. https://doi.org/10.1001/archpsyc.55.5.393

[40] Elbogen, E.B. and Johnson, S.C. (2009) The Intricate Link between Violence and Mental Disorder: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Archives of General Psychiatry*, **66**, 152-161. https://doi.org/10.1001/archgenpsychiatry.2008.537

[41] Kay, S.R., Fiszbein, A. and Opler, L.A. (1987) The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, **13**, 261-276. https://doi.org/10.1093/schbul/13.2.261

[42] Montgomery, S.A. and Åsberg, M. (1979) A New Depression Scale Designed to Be Sensitive to Change. *British Journal of Psychiatry*, **134**, 382-389. https://doi.org/10.1192/bjp.134.4.382

[43] Immanuel, S., Teferra, M.N., Baumert, M. and Bidargaddi, N. (2023) Heart Rate Variability for Evaluating Psychological Stress Changes in Healthy Adults: A Scoping Re-

view. *Neuropsychobiology*, **82**, 187-202. https://doi.org/10.1159/000530376

[44] Dawson, M.E., Schell, A.M. and Filion, D.L. (2007) The Electrodermal System. *Handbook of Psychophysiology*, **2**, 200-223.

[45] Kudielka, B.M., Hellhammer, D.H. and Wüst, S. (2009) Why Do We Respond So Differently? Reviewing Determinants of Human Salivary Cortisol Responses to Challenge. *Psychoneuroendocrinology*, **34**, 2-18. https://doi.org/10.1016/j.psyneuen.2008.10.004

[46] Nijman, H.L.I., Palmstierna, T., Almvik, R. and Stolker, J.J. (2005) Fifteen Years of Research with the Staff Observation Aggression Scale: A Review. *Acta Psychiatrica Scandinavica*, **111**, 12-21. https://doi.org/10.1111/j.1600-0447.2004.00417.x

[47] Forrest, S., Eatough, V. and Shevlin, M. (2004) Measuring Adult Indirect Aggression: The Development and Psychometric Assessment of the Indirect Aggression Scales. *Aggressive Behavior*, **31**, 84-97. https://doi.org/10.1002/ab.20074

[48] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[49] Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors). *The Annals of Statistics*, **28**, 337-407. https://doi.org/10.1214/aos/1016218223

[50] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/a:1022627411411

[51] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[52] Zhang, Z. and Jung, C. (2021) GBDT-MO: Gradient-Boosted Decision Trees for Multiple Outputs. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 3156-3167. https://doi.org/10.1109/tnnls.2020.3009776

[53] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd Edition, Wiley. https://doi.org/10.1002/9781118548387

[54] Chang, C. and Lin, C. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27. https://doi.org/10.1145/1961189.1961199

[55] Friedman, J.H. (1998) Data Mining and Statistics: What's the Connection? *Computing Science and Statistics*, **29**, 3-9.

[56] Arlot, S. and Celisse, A. (2010) A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, **4**, 40-79. https://doi.org/10.1214/09-ss054

[57] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[58] Van Rijsbergen, C. (1979) Information Retrieval: Theory and Practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, Butterworth-Heinemann, Vol. 79, 1-14.

[59] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., *et al.* (2010) Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*, **21**, 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2

[60] Hesterberg, T. (2011) Bootstrap. *WIREs Computational Statistics*, **3**, 497-526. https://doi.org/10.1002/wics.182

[61] Adler, A.I. and Painsky, A. (2022) Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy*, **24**, 687. https://doi.org/10.3390/e24050687

[62] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[63] Murphy, A.H. (1973) A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, **12**, 595-600. https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2

[64] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-845. https://doi.org/10.2307/2531595

[65] Austin, P.C. and Steyerberg, E.W. (2014) Events per Variable (EPV) and the Relative Performance of Different Strategies for Estimating the Out-of-Sample Validity of Logistic Regression Models. *Statistical Methods in Medical Research*, **26**, 796-808. https://doi.org/10.1177/0962280214558972

[66] Yang, M., Wong, S.C.P. and Coid, J. (2010) The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools. *Psychological Bulletin*, **136**, 740-767. https://doi.org/10.1037/a0020473

[67] Singh, J.P., Desmarais, S.L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., *et al.* (2014) International Perspectives on the Practical Application of Violence Risk Assessment: A Global Survey of 44 Countries. *International Journal of Forensic Mental Health*, **13**, 193-206. https://doi.org/10.1080/14999013.2014.922141

[68] Kendler, K.S., Ohlsson, H., Morris, N.A., Sundquist, J. and Sundquist, K. (2014) A Swedish Population-Based Study of the Mechanisms of Parent-Offspring Transmission of Criminal Behavior. *Psychological Medicine*, **45**, 1093-1102. https://doi.org/10.1017/s0033291714002268

[69] Harris, A. and Lurigio, A.J. (2007) Mental Illness and Violence: A Brief Review of Research and Assessment Strategies. *Aggression and Violent Behavior*, **12**, 542-551. https://doi.org/10.1016/j.avb.2007.02.008

[70] Regier, D.A. (1990) Comorbidity of Mental Disorders with Alcohol and Other Drug Abuse. *JAMA*, **264**, 2511. https://doi.org/10.1001/jama.1990.03450190043026

[71] Beauchaine, T.P. and Thayer, J.F. (2015) Heart Rate Variability as a Transdiagnostic Biomarker of Psychopathology. *International Journal of Psychophysiology*, **98**, 338-350. https://doi.org/10.1016/j.ijpsycho.2015.08.004

[72] Thayer, J.F., Hansen, A.L., Saus-Rose, E. and Johnsen, B.H. (2009) Heart Rate Variability, Prefrontal Neural Function, and Cognitive Performance: The Neurovisceral Integration Perspective on Self-Regulation, Adaptation, and Health. *Annals of Behavioral Medicine*, **37**, 141-153. https://doi.org/10.1007/s12160-009-9101-z

[73] Price, O. and Baker, J. (2012) Key Components of De-Escalation Techniques: A Thematic Synthesis. *International Journal of Mental Health Nursing*, **21**, 310-319. https://doi.org/10.1111/j.1447-0349.2011.00793.x

[74] Riddell, J., Tran, A., Bengiamin, R., Hendey, G.W. and Armenian, P. (2017) Ketamine as a First-Line Treatment for Severely Agitated Emergency Department Patients. *The American Journal of Emergency Medicine*, **35**, 1000-1004. https://doi.org/10.1016/j.ajem.2017.02.026

[75] Sailas, E.E. and Fenton, M. (2000) Seclusion and Restraint for People with Serious Mental Illnesses. *Cochrane Database of Systematic Reviews*, **2012**, CD001163. https://doi.org/10.1002/14651858.cd001163

[76] Huckshorn, K.A. (2006) Re-Designing State Mental Health Policy to Prevent the Use of Seclusion and Restraint. *Administration and Policy in Mental Health and Mental Health Services Research*, **33**, 482-491. https://doi.org/10.1007/s10488-005-0011-5

[77] Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., *et al*. (2012) Shared Decision Making: A Model for Clinical Practice. *Journal of General Internal Medicine*, **27**, 1361-1367. https://doi.org/10.1007/s11606-012-2077-6

[78] Stiggelbout, A.M., Pieterse, A.H. and De Haes, J.C.J.M. (2015) Shared Decision Making: Concepts, Evidence, and Practice. *Patient Education and Counseling*, **98**, 1172-1179. https://doi.org/10.1016/j.pec.2015.06.022

[79] Corrigan, P.W. and Watson, A.C. (2002) The Paradox of Self-Stigma and Mental Illness. *Clinical Psychology*: *Science and Practice*, **9**, 35-53. https://doi.org/10.1093/clipsy.9.1.35

[80] Thornicroft, G., Rose, D. and Kassam, A. (2007) Discrimination in Health Care against People with Mental Illness. *International Review of Psychiatry*, **19**, 113-122. https://doi.org/10.1080/09540260701278937

[81] Dash, S., Syed, Y.A. and Khan, M.R. (2022) Understanding the Role of the Gut Microbiome in Brain Development and Its Association with Neurodevelopmental Psychiatric Disorders. *Frontiers in Cell and Developmental Biology*, **10**, Article ID: 880544. https://doi.org/10.3389/fcell.2022.880544

[82] Trevey, C. (2010) Prisoners of the Mind?: The Inappropriateness of Comparing the Involuntarily Committed Mentally Ill to Pretrial Detainees in Fourth Amendment Analyses. *The University of Pennsylvania Journal of Constitutional Law*, **13**, 1435.

[83] Arbach-Lucioni, K., Andrés-Pueyo, A., Pomarol-Clotet, E. and Gomar-Soñes, J. (2011) Predicting Violence in Psychiatric Inpatients: A Prospective Study with the HCR-20 Violence Risk Assessment Scheme. *Journal of Forensic Psychiatry & Psychology*, **22**, 203-222. https://doi.org/10.1080/14789949.2010.530290

[84] Belsher, B.E., Smolenski, D.J., Pruitt, L.D., Bush, N.E., Beech, E.H., Workman, D.E., *et al*. (2019) Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. *JAMA Psychiatry*, **76**, 642-651. https://doi.org/10.1001/jamapsychiatry.2019.0174

[85] Ursano, R.J., Kessler, R.C., Stein, M.B., Naifeh, J.A., Aliaga, P.A., Fullerton, C.S., *et al*. (2015) Suicide Attempts in the US Army during the Wars in Afghanistan and Iraq, 2004 to 2009. *JAMA Psychiatry*, **72**, 917-926. https://doi.org/10.1001/jamapsychiatry.2015.0987

[86] Fazel, S., Wolf, A., Palm, C. and Lichtenstein, P. (2014) Violent Crime, Suicide, and Premature Mortality in Patients with Schizophrenia and Related Disorders: A 38-Year Total Population Study in Sweden. *The Lancet Psychiatry*, **1**, 44-54. https://doi.org/10.1016/s2215-0366(14)70223-8

[87] Wallace, C., Mullen, P.E. and Burgess, P. (2004) Criminal Offending in Schizophrenia over a 25-Year Period Marked by Deinstitutionalization and Increasing Prevalence of Comorbid Substance Use Disorders. *American Journal of Psychiatry*, **161**, 716-727. https://doi.org/10.1176/appi.ajp.161.4.716

[88] Chow, W.S. and Priebe, S. (2013) Understanding Psychiatric Institutionalization: A Conceptual Review. *BMC Psychiatry*, **13**, Article No. 169. https://doi.org/10.1186/1471-244x-13-169

[89] Sashidharan, S.P., Mezzina, R. and Puras, D. (2019) Reducing Coercion in Mental Healthcare. *Epidemiology and Psychiatric Sciences*, **28**, 605-612. https://doi.org/10.1017/s2045796019000350

[90] Bedoya, A.D., Futoma, J., Clement, M.E., Corey, K., Brajer, N., Lin, A., *et al*. (2020)

Machine Learning for Early Detection of Sepsis: An Internal and Temporal Validation Study. *JAMIA Open*, **3**, 252-260. https://doi.org/10.1093/jamiaopen/ooaa006

[91] Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., *et al.* (2020) A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Proceedings of the* 2020 *CHI Conference on Human Factors in Computing Systems*, Honolulu, 25-30 April 2020, 1-12. https://doi.org/10.1145/3313831.3376718

[92] Steyerberg, E.W. and Harrell, F.E. (2016) Prediction Models Need Appropriate Internal, Internal-External, and External Validation. *Journal of Clinical Epidemiology*, **69**, 245-247. https://doi.org/10.1016/j.jclinepi.2015.04.005

[93] Riley, R.D., Ensor, J., Snell, K.I.E., Debray, T.P.A., Altman, D.G., Moons, K.G.M., *et al.* (2016) External Validation of Clinical Prediction Models Using Big Datasets from E-Health Records or IPD Meta-Analysis: Opportunities and Challenges. *BMJ*, **353**, i3140. https://doi.org/10.1136/bmj.i3140

[94] Torous, J., Nicholas, J., Larsen, M.E., Firth, J. and Christensen, H. (2018) Clinical Review of User Engagement with Mental Health Smartphone Apps: Evidence, Theory and Improvements. *Evidence Based Mental Health*, **21**, 116-119. https://doi.org/10.1136/eb-2018-102891

[95] Mohr, D.C., Zhang, M. and Schueller, S.M. (2017) Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, **13**, 23-47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

[96] Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., *et al.* (2009) Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, **338**, b2393-b2393. https://doi.org/10.1136/bmj.b2393

[97] van Buuren, S. (2018) Flexible Imputation of Missing Data. 2nd Edition, CRC Press.

[98] Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B. and Richardson, W.S. (1996) Evidence Based Medicine: What It Is and What It Isn't. *BMJ*, **312**, 71-72. https://doi.org/10.1136/bmj.312.7023.71

[99] Glasziou, P., Chalmers, I., Rawlins, M. and McCulloch, P. (2007) When Are Randomised Trials Unnecessary? Picking Signal from Noise. *BMJ*, **334**, 349-351. https://doi.org/10.1136/bmj.39070.527986.68

[100] Grisso, T., Davis, J., Vesselinov, R., Appelbaum, P.S. and Monahan, J. (2000) Violent Thoughts and Violent Behavior Following Hospitalization for Mental Disorder. *Journal of Consulting and Clinical Psychology*, **68**, 388-398. https://doi.org/10.1037/0022-006x.68.3.388

[101] Tardiff, K. (1998) Unusual Diagnoses among Violent Patients. *Psychiatric Clinics of North America*, **21**, 567-576. https://doi.org/10.1016/s0193-953x(05)70023-x

[102] Fogel, A.L. and Kvedar, J.C. (2018) Artificial Intelligence Powers Digital Medicine. *NPJ Digital Medicine*, **1**, Article No. 5. https://doi.org/10.1038/s41746-017-0012-2

[103] Obermeyer, Z. and Emanuel, E.J. (2016) Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, **375**, 1216-1219. https://doi.org/10.1056/nejmp1606181

[104] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C. and Faisal, A.A. (2018) The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care. *Nature Medicine*, **24**, 1716-1720. https://doi.org/10.1038/s41591-018-0213-5

[105] Nemati, S., Holder, A., Razmi, F., Stanley, M.D., Clifford, G.D. and Buchman, T.G. (2018) An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*, **46**, 547-553. https://doi.org/10.1097/ccm.0000000000002936

[106] Collins, G.S. and Moons, K.G.M. (2019) Reporting of Artificial Intelligence Prediction Models. *The Lancet*, **393**, 1577-1579. https://doi.org/10.1016/s0140-6736(19)30037-6

[107] Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., *et al.* (2016) Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of Medical Internet Research*, **18**, e323. https://doi.org/10.2196/jmir.5870

[108] Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., *et al.* (2017) Risk Factors for Suicidal Thoughts and Behaviors: A Meta-Analysis of 50 Years of Research. *Psychological Bulletin*, **143**, 187-232. https://doi.org/10.1037/bul0000084

[109] Olfson, M., Wall, M., Wang, S., Crystal, S., Liu, S., Gerhard, T., *et al.* (2016) Short-Term Suicide Risk after Psychiatric Hospital Discharge. *JAMA Psychiatry*, **73**, 1119-1126. https://doi.org/10.1001/jamapsychiatry.2016.2035

[110] Delanerolle, G., Bouchareb, Y., Shetty, S., Cavalini, H. and Phiri, P. (2025) A Pilot Study Using Natural Language Processing to Explore Textual Electronic Mental Healthcare Data. *Informatics*, **12**, Article No. 28. https://doi.org/10.3390/informatics12010028

[111] Martin-Sanchez, F., Iakovidis, I., Nørager, S., Maojo, V., de Groen, P., Van der Lei, J., *et al.* (2004) Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Health Care. *Journal of Biomedical Informatics*, **37**, 30-42. https://doi.org/10.1016/j.jbi.2003.09.003

[112] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., *et al.* (2020) The Future of Digital Health with Federated Learning. *NPJ Digital Medicine*, **3**, 1-7. https://doi.org/10.1038/s41746-020-00323-1

[113] Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., *et al.* (2020) Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, **10**, Article No. 12598. https://doi.org/10.1038/s41598-020-69250-1

[114] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[115] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G. (2019) XAI—Explainable Artificial Intelligence. *Science Robotics*, **4**, eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[116] Price, W.N. and Cohen, I.G. (2019) Privacy in the Age of Medical Big Data. *Nature Medicine*, **25**, 37-43. https://doi.org/10.1038/s41591-018-0272-7

[117] Machireddy, J.R. (2023) Harnessing AI and Data Analytics for Smarter Healthcare Solutions. *International Journal of Science and Research Archive*, **8**, 785-798. https://doi.org/10.30574/ijsra.2023.8.2.0105