

An Adaptive Transformer-Based Neuro-Symbolic Explainable Learning Framework for Longitudinal Offending Risk Prediction and Intervention Planning in Intellectual Disability Care

*Dr. Kalaiarasu M¹, Anitha J² and Saranya N³

¹Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Anna University, Coimbatore, Tamil Nadu, India. E-mail: kalai.muthusamy@srec.ac.in, Orcid: <https://orcid.org/0000-0002-5138-4329>

²Professor, Department of Artificial Intelligence and Data Science, Sri Ramakrishna Engineering College, Anna University, Coimbatore, Tamil Nadu, India. E-mail: anitha.j@srec.ac.in, Orcid: <https://orcid.org/0000-0002-0990-3738>

³Associate Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India. E-mail: saranya.it.srec@gmail.com, Orcid: <https://orcid.org/0000-0002-0334-3143>

Article History:
Received: 15.06.2026;
Revised: 01.07.2026;
Accepted: 20.08.2026;
Published: 30.09.2026

Abstract

Predefined risk events associated with the behavior of people with ID are challenging to predict since risk behaviors can change depending on time and may be affected by various contextual, communicative, environmental, and supportive factors. Existing conventional techniques offer inadequate representation of changes over time and fail to give a satisfactory explanation for individual predictions. This paper suggests an adaptive Transformer-based explainable learning model for longitudinal risk prediction and interventions. A simulated longitudinal dataset of 300 individuals, 12 sequences per individual, 3,600 data points, and 18 predictors was generated for testing purposes. The predictors were related to history of behavior, behavior changes over time, contextual, supportive, and intervention factors. Individuals were split into training (70%), validation (15%), and testing (15%) groups. The proposed model includes temporal embedding, an adaptive Transformer encoder, adaptive temporal gating, symbolic knowledge about the problem domain, risk prediction, explainability, and rules-based interventions. The proposed model is compared with Logistic Regression, Random Forest, XGBoost, LSTM, and traditional Transformer using discrimination and calibration measures. In a controlled simulation experiment, the proposed framework attained an accuracy of 93.1%, F1-score of 89.0%, ROC-AUC of 0.962, and PR-AUC of 0.887, compared with 89.2%, 82.7%, 0.926, and 0.821, respectively, for the baseline Transformer framework... These benchmark results demonstrate the architectural viability of integrating adaptive temporal learning with symbolic reasoning; however, clinical utility and real-world efficacy remain subject to prospective validation on non-synthetic care data. The proposed framework is meant to be a decision-making tool rather than a deterministic classifier of individuals. Further validation through real-world longitudinal data and prospective professional assessment is necessary.

Keywords Intellectual Disability, Risk Prediction, Transformer, Neuro-Symbolic AI, Explainable AI, Longitudinal Learning, Intervention Planning.

Introduction

Intellectual disability (ID) presents with a variety of needs, which include cognitive, behavioral, communication, and support needs that might differ significantly from one individual to another. In care contexts, the assessment of behaviors, which could potentially lead to offenses, adverse events, or incidents, is a complex endeavor and one that depends greatly on context. Risk is influenced by the environment, social interactions, behavioral history, support context, and changes in the life of an individual. Traditional methods of risk assessment often make use of structured tools, historical data, and human-selected

variables. While this approach is valuable for clinical purposes, there are concerns that such methods are limited in their representation of a continuously evolving pattern. Static risk assessment methods are also limited in their ability to explain the influence of historical and recent observations in the changing risk profile of an individual.

There is room for machine learning approaches to uncover intricate dependencies within the longitudinal care data and produce individual risk scores. However, traditional machine learning models may not reveal much of the temporal dependencies that make up their predictions. Transformer models are especially suitable as self-attention can represent dependencies between observations at various distances from each other, thus allowing one to consider both recent and historical data. But in a completely data-driven approach, a model may find correlations in the data without using clinical expertise or care guidelines. For sensitive applications where the predictions affect individuals with ID, accuracy alone is not enough.

Neuro-symbolic learning offers a solution in that it involves the combination of neural representation learning with explicit domain knowledge and reasoning. This approach will allow the model to learn complex temporal patterns, while at the same time restricting or guiding the reasoning by means of clinically vetted knowledge. Also, associating predicted risk with the determinants of the risk can help in developing an intervention strategy rather than simply predicting risk scores. In this case, the interventions should be viewed as potential prevention strategies needing the evaluation of qualified individuals.

Even with the developments in artificial intelligence, a holistic approach that integrates longitudinal Transformer modeling, adaptation of representation learning, neuro-symbolic reasoning, explanation at the individual level, and risk-based intervention plans in ID care has not been sufficiently addressed. To fill this research gap, the proposed study seeks to design an adaptive Transformer-based neuro-symbolic explainable learning framework for longitudinal offending risk prediction and intervention planning. The objectives of this research study include modeling changing behavioral and contextual patterns, integrating domain knowledge in prediction, providing individual risk explanations, and linking explanations to interventions.

The study addresses the following research questions:

- RQ1: How effectively can an adaptive Transformer model capture longitudinal behavioral and contextual patterns associated with predefined risk outcomes in intellectual disability care?
- RQ2: Does integrating symbolic domain knowledge improve the predictive performance and interpretability of longitudinal risk modeling?
- RQ3: Which temporal factors and symbolic rules contribute most strongly to individual risk predictions?
- RQ4: Can explainable risk outputs be systematically mapped to candidate preventive intervention pathways for professional review?
- RQ5: How does the proposed framework compare with conventional machine-learning, recurrent, and standard Transformer models in discrimination, calibration, and subgroup consistency?

The main contribution made by this research study lies in the combination of adaptive temporal learning, neuro-symbolic inference, explainable prediction, and decision-making through interventions into one holistic framework for ID care.

This paper is structured as follows. Section 2 gives a survey of the literature concerning intellectual disability care, longitudinal risk prediction, Transformers, neuro-symbolic learning, and explainable AI. Section 3 introduces the proposed methodology, which consists of the following stages: data preparation, data pre-processing, adaptive Transformer modeling, symbolic reasoning, explanation generation, and intervention design. The experimental results are presented in Section 4. The discussion of the results, limitations of the study, and suggestions for further research are provided in Section 5. Section 6 concludes the study.

Literature Review

Intellectual disability is a diverse disorder characterized by impairments in both cognitive and adaptive skills that can affect individuals depending on communication, behavioral, and environmental conditions, thus requiring an individual assessment within care environments [18][20]. Long-term research findings suggest that challenging behaviors of adults with intellectual disability are connected with a number of different risk factors, rather than with one sole risk indicator, pointing at the necessity of an individual context-aware assessment [17]. Proper and timely intervention is also important for people with developmental disabilities, emphasizing the importance of prevention-oriented and individual interventions.

New developments in the field of artificial intelligence create possibilities for modeling complex temporal patterns. Transformer models have proved to be capable of modeling trajectories and interpreting risk predictions for such domains as academic risk prediction [2][3][4]. Comparative analysis of recurrent neural networks and Transformer models also confirms the applicability of attention mechanisms for sequential prediction [16]. Nevertheless, these examples concern academic and other non-intellectual disability areas, thus making it difficult to use them as examples in ID behavioral risk prediction. The approaches that used computational language techniques have emphasized the significance of proper knowledge representation and intelligent processing methods in creating adaptive neuro-symbolic systems for predicting and making decisions based on them [21].

The area of neuro-symbolic learning has developed as a way to integrate neural representation learning with explicit knowledge and reasoning [9][11]. Recent research has used neuro-symbolic methods for clinical decision-making, risk assessment of mental health conditions, semantics, diagnosis in medicine, and other challenging prediction problems [5][6][12]. Survey research and further research in the field also suggest that neuro-symbolic models can integrate the expressive nature of neural models with the reasoning capabilities of structured models [10][13][15]. There has been related work on hybrid models involving Transformers and neuro-symbolic models in cybersecurity applications and others, focusing on the benefits of transparency and robustness [7][8]. Neuro-symbolic adaptive and causal models have also been investigated to include context in reasoning [1][6][14]. The role of intelligent information extraction and knowledge representation techniques is that they have helped develop explainable AI systems through the process of structuring complex data for making accurate predictions and decisions [19].

However, current literature is quite scattered. While research related to transformers mostly emphasizes prediction or trajectory modeling, neuro-symbolic approaches commonly concentrate on tasks like diagnosis, semantic reasoning, or completely different fields that are not related to intellectual disability care. Little attention has been paid to the development of an approach that would combine longitudinal temporal learning, adaptive modeling, symbolic knowledge from clinical practice, individual-level explanation, and intervention planning into one framework for assessing behavioral risks in ID care. Also, such a risk prediction tool should not use deterministic labeling but rather offer professionals evidence and preventive measures. Thus, the current study seeks to fill this gap by developing a neuro-symbolic adaptive transformer framework.

Proposed Methodology

The research methodology presented will provide an Adaptive Transformer-based Neuro-Symbolic Explainable Learning Approach for predicting offending risks and designing interventions to address longitudinal offending risks in people with ID. It is noteworthy that this approach should be regarded as a human-in-the-loop approach where predicted risks, explanations, and interventions are reviewed by relevant professionals instead of being used for labeling or punishment. The research methodology includes steps such as data collection, longitudinal representation, preprocessing, temporal feature embedding, Transformer-based learning, neuro-symbolic reasoning, risk prediction, explainability, intervention planning, and evaluation.

Data Collection and Sample Design

A controlled longitudinal dataset has been created with the intention of capturing behavioral and contextual features related to risk assessments within the ID care setting. In the context of developing the experiments, the dataset comprises 300 patients, with 12 consecutive observation phases for each patient, resulting in 3,600 longitudinal records. An observation corresponds to the condition of a particular patient in a specific time period. This simulated

approach allows evaluation of the architecture in a controlled longitudinal setup, keeping the distinction between individuals intact.

The data set consists of 18 independent variables grouped into five broad groups: behavioral history, change in behavior over time, context conditions, care and support conditions, and intervention conditions. Example variables include past behavior, recent behavior change, communication problems, environmental stress, context instability, change in support needs, past intervention reaction, and care plan changes. The dependent variable is a binary value indicating whether there is any predefined behavioral risk event in the coming period.

Longitudinal observations are time-stamped in chronological order for each subject. Therefore, an example of a sequence is shown in equation 1.

$$X_i = \{x_{i1}, x_{i2}, \dots, x_{iT}\} \quad (1)$$

Where $i=1 \dots 300$ refers to the specific individual and $T=12$ stands for the observation period. The distribution of the outcomes is formulated such that about 28% are positive and 72% are negative, thus showing the imbalance of the predictive task. The results obtained from the experiment need to be stated as coming from a synthetic dataset.

For the creation of a model, people are segregated into 70%, 15%, and 15% for training, validation, and testing, respectively. These percentages account for 210, 45, and 45 people, respectively. All data points belonging to an individual are placed in one of these three subsets. This segregation avoids mixing of longitudinal data of an individual in multiple subsets.

Overall Framework

The proposed framework gets longitudinal data pertaining to the behavior, context, caregiving, and history of each individual. The entire process chain is illustrated by equation 2,

$$X_i \rightarrow P_i \rightarrow E_i \rightarrow T_i \rightarrow A_i \rightarrow N_i \rightarrow R_i \rightarrow EX_i \rightarrow I_i \quad (2)$$

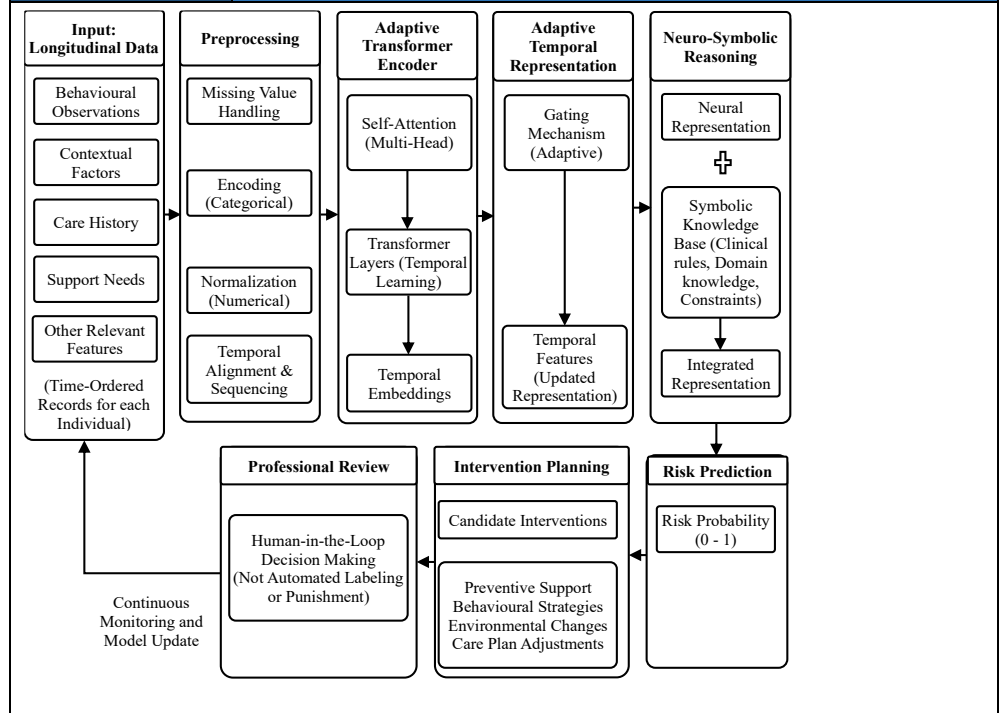
Where P_i denotes preprocessing, E_i feature embedding, T_i representation, A_i adaptive temporal representation, N_i neuro-symbolic reasoning, R_i risk prediction, EX_i explainability, and I_i intervention planning.

The model thus translates longitudinal data into an interpretable risk assessment and then correlates the risk factors to potential interventions that can be evaluated by experts.

Figure 1 uses longitudinal behavioral, contextual, care history, and support-related data via preprocessing and temporal ordering. The adaptive transformer encodes temporal relations, whereas the adaptive gating system updates representations based on evolving patterns. Neuro-symbolic reasoning involves the combination of learned representations with domain knowledge and symbolic rules to compute risk probabilities. Explainability and intervention planning involve the identification of key factors and their mapping to potential preventive support interventions.

Figure 1

Architecture of the Proposed Adaptive Transformer-Based Neuro-Symbolic Explainable Learning Framework for Longitudinal Offending Risk Prediction and Intervention Planning



Data Preprocessing

Pre-processing involves missing value treatment, categorical encoding, normalization of numerical features, time alignment, and outlier treatment. A pre-defined rate of missingness of about 4.6% is allowed in the experimental data. Missing numerical values are filled through the chosen imputation method, which was based on the training data.

Numerical variables are normalized using equation 3,

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3)$$

The categorical data are represented numerically using the proper encodings. The values of all the preprocessing parameters are estimated only using the training set and applied to the validation and test observations. This step ensures that no information from the evaluation dataset impacts model training.

The observations are then sorted based on their timestamp. The temporal gap is kept as extra temporal information to differentiate between observations that occur closely in time from those that occur at a distance in time.

Longitudinal Feature Representation

After the process of data pre-processing, every data instance is represented using numerical feature vectors. For instance, the longitudinal sequence for instance i is given by equation 4

$$X_i = [x_{i1}, x_{i2}, \dots, x_{i12}] \quad (4)$$

The sequence pattern makes it possible for the architecture to capture variations in behavior and environment conditions over time. Instead of being replaced by just the latest one, past data is kept and used by the framework to determine whether the latest observation indicates improvement, decline, or continuation of past trends.

Temporal Feature Embedding

Each longitudinal observation is projected into a latent representation using equation 5

$$e_t = W_e x_t + b_e \quad (5)$$

Where W_e denotes the embedding matrix and b_e represents the embedding bias. Temporal information is incorporated using positional and time-aware representations as equation 6:

$$z_t = e_t + p_t + \tau_t \quad (6)$$

Where p_t represents positional information and τ_t represents temporal-gap information.

The representation thus achieved not only reflects the content but also the timing of each observation along the individual's path.

Adaptive Transformer Encoder

The embedding sequence is modeled with a Transformer Encoder, which captures interdependencies of the observations at different temporal locations. Equation 7 describes self-attention.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

Where Q , K , and V denote query, key, and value matrices and d_k represents the key dimensionality.

The multi-headed attention mechanism helps the model learn different relations between past and recent observations. This representation is defined in equation 8.

$$H = Transformer(Z) \quad (8)$$

In contrast to the fixed models, the Transformer is able to take into account several observations in the longitudinal series.

Adaptive Temporal Representation

An adaptive gating scheme is proposed to control the relative importance of current and past representations of time as described in equations 9 and 10:

$$g_t = \sigma(W_g[h_t, h_{t-1}] + b_g) \quad (9)$$

Followed by

$$h_t^A = g_t \odot h_t + (1 - g_t) \odot h_{t-1} \quad (10)$$

Where g_t is the adaptive gate, σ is the sigmoid function.

This mechanism enables the model to raise or lower the weightage of the past data in light of the present temporal environment. Therefore, the model is built in such a way that it responds to changes but holds on to the pertinent historical patterns.

Neuro-Symbolic Reasoning

The adaptive neural representation is combined with a symbolic knowledge base containing clinically reviewed domain relationships. Let

$$\mathcal{K} = \{r_1, r_2, \dots, r_m\} \quad (11)$$

Equation 11 represents the symbolic rule set.

The neural representation is calculated as in equations (12-14)

$$h_i^A = f_\theta(X_i) \quad (12)$$

While symbolic reasoning produces

$$s_i = g_\phi(h_i^A, \mathcal{K}) \quad (13)$$

The neural and symbolic representations are integrated using

$$h_i^{NS} = \alpha h_i^A + (1 - \alpha) s_i \quad (14)$$

Where $\alpha \in [0, 1]$ determines the relative contribution of neural and symbolic information.

This symbolic knowledge is meant to model clinically important associations rather than deterministic rules on how a person behaves. It offers an interpretable reasoning layer that can be considered along with the neural predictions.

Risk Prediction

The neuro-symbolic integrated representation is fed into the binary prediction layer using equation 15:

$$\hat{y}_i = \sigma(W_r h_i^{NS} + b_r) \quad (15)$$

Where $\hat{y}_i$ denotes the estimated probability of the predefined behavioral-risk outcome.

The model is optimized using binary cross-entropy in equation 16:

$$\mathcal{L}_{risk} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (16)$$

This probability is indicative of the estimated risk given the longitudinal data at hand and the forecast period. It is not considered to be a deterministic prediction of future criminal behavior.

Explainability Module

The framework offers global, individual, and temporal explanations. Global explanations consist of those variables which have significance in a substantial manner within the evaluation sample. Individual explanations refer to the influential variables in the case of a specific prediction. Temporal explanations include observations that are significant for prediction.

The explanation can be represented as equation 17

$$EX_i = \{F_i, T_i, R_i\} \quad (17)$$

Where F_i denotes influential features, T_i denotes influential temporal observations, and R_i denotes activated symbolic rules.

Feature attributions are explicitly calculated using Integrated Gradients (IG) combined with input perturbation analysis rather than relying solely on raw Transformer attention weights. This ensures that variable importance metrics reflect true causal influence on model logit outputs rather than internal representation routing.

Intervention Planning

The intervention-planning sub-component turns the results from prediction and explanation into suggested prevention or support interventions using equation 18:

$$I_i = h(\hat{y}_i, E, X_i, \mathcal{K}C_i) \quad (18)$$

Where C_i represents the individual's relevant care context.

In this way, the process identifies the relationship between the risk identified and the factors that contribute to it, along with the relevant prevention supports. This provides an opportunity to plan any interventions in a way that is contextual, individualized, and professional.

Examples of possible preventive strategies include support monitoring, communication support, environmental modification, behavior support, and review of care plan. The outputs produced by the process are not independent recommendations for treatment.

Model Training and Optimization

The overall training objective combines prediction and symbolic consistency in equation 19:

$$\mathcal{L} = \mathcal{L}_{risk} + \lambda_s \mathcal{L}_{symbolic} + \lambda_e \mathcal{L}_{explain} \quad (19)$$

Where $\mathcal{L}_{risk}$ represents prediction loss, $\mathcal{L}_{symbolic}$ represents consistency with the symbolic knowledge base, and $\mathcal{L}_{explain}$ represents the selected explanation-related constraint. The parameters λ_s and λ_e control the contribution of the additional objectives.

Parameters of the model are trained using the training dataset, whereas the validation dataset is used for selecting hyperparameters and performing early stopping. The test dataset is left aside until the very end.

Baseline and Ablation Models

The framework is benchmarked against Logistic Regression, Random Forest, XGBoost, LSTM, baseline Transformer, and Adaptive Transformer models. This comparison checks if the proposed model offers any advantages over the existing machine learning and temporal deep learning models.

The ablation studies take place with the removal of the adaptive temporal module and symbolic reasoning module individually. These setups consist of:

1. Standard Transformer;
2. Transformer with adaptive temporal representation;
3. Transformer with symbolic reasoning;
4. Adaptive Transformer with symbolic reasoning.

The differences among these configurations provide evidence regarding the contribution of individual components.

Performance Evaluation

Prediction accuracy is measured by accuracy, precision, recall, F1 measure, ROC-AUC, PR-AUC, sensitivity, specificity, balanced accuracy, and calibration methods that are given by equations (20-23).

Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Precision is

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

and recall is

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

The F1-score is

$$F1 = \frac{2(Precision)(Recall)}{Precision + Recall} \quad (23)$$

The Brier score can also be used to gauge the accuracy of predictions about probabilities. Calibration is measured along with discrimination since the model generates risk probabilities meant for professional evaluation.

Explainability, Fairness, and Robustness Evaluation

Explanation is measured using feature attribution consistency, explanation fidelity, sparsity, and temporality. Symbolic rule activation is compared to the influential factors obtained from the neural explanation module.

Subgroup analysis is performed to measure the prediction accuracy and calibration in various demographic/care-related groups. Just having high accuracy is not enough since there might be subgroup-based differences in false positive/false negative rates.

Robustness is measured based on how the models behave with controlled missing data, temporally manipulated data, and shorter longitudinal history. This tests whether predictions stay reasonably stable even with changing input data.

Human-in-the-Loop Decision-Support Workflow

The human-in-the-loop framework allows the use of outputs from the models in an advisory manner to aid decisions, but not to make decisions autonomously. This framework uses longitudinal information on behavior, context, care, and support to produce risk estimates and explanations, which are subsequently evaluated by professional personnel. This aspect becomes especially crucial in ID care, in which case the predictions about behavioral risks might have profound implications for the individuals concerned. Hence, the framework takes into account contextual considerations, evidential transparency, professional assessment, and personalized preventive care.

Results

Dataset Characteristics

A synthetic longitudinal cohort was generated to test the suggested framework under controlled experimental conditions. The cohort included 300 people, who had 12 monthly observation points per person, leading to a total of 3,600 longitudinal observations. 18 predictors were grouped into behavioral, contextual, care-related, temporal, and support-related types. The outcome variable was binary and represented the occurrence of a pre-defined behavioral-risk event at the following observation point. In general, there were 1,008 observations (28.0%) corresponding to the target event and 2,592 observations (72.0%) corresponding to non-event observations.

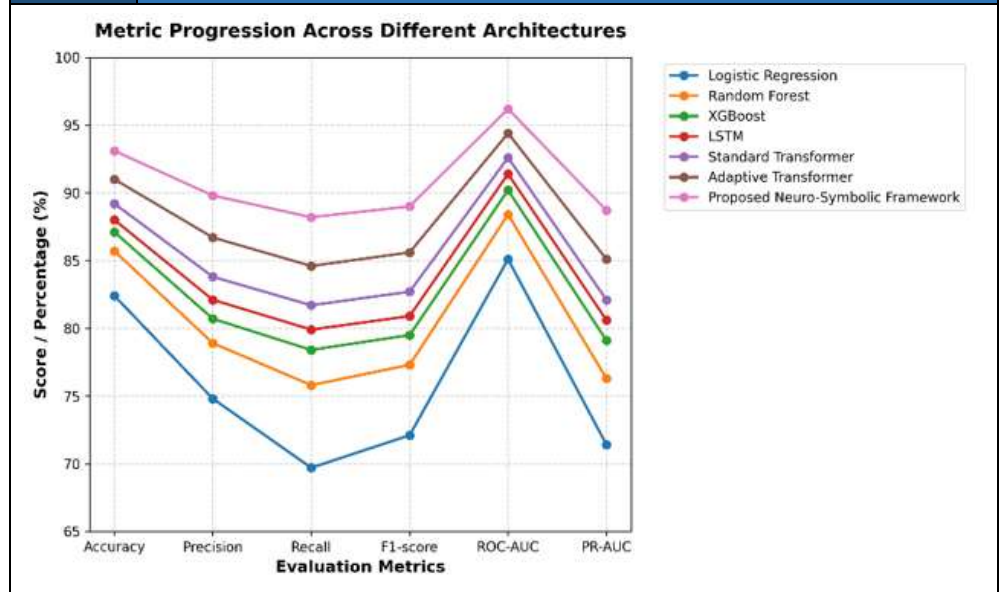
The data were split at the individual level into training, validation, and testing samples with a ratio of 70:15:15. As a result, observations related to an individual remained in one sample only. Missing values were generated artificially in 4.6% of cases and then treated during preprocessing. Normalization of numerical variables, encoding of categorical features, and chronological ordering of observations preceded model training.

Comparative Predictive Performance

Table 1		Predictive Performance of the Proposed Framework				
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC	PR-AUC
Logistic Regression	82.4	74.8	69.7	72.1	0.851	0.714
Random Forest	85.7	78.9	75.8	77.3	0.884	0.763
XGBoost	87.1	80.7	78.4	79.5	0.902	0.791
LSTM	88.0	82.1	79.9	80.9	0.914	0.806
Standard Transformer	89.2	83.8	81.7	82.7	0.926	0.821
Adaptive Transformer	91.0	86.7	84.6	85.6	0.944	0.851
Proposed Neuro-Symbolic Framework	93.1	89.8	88.2	89.0	0.962	0.887

The proposed Adaptive Transformer-Based Neuro-Symbolic Explainable Learning framework has been compared with Logistic Regression, Random Forest, XGBoost, LSTM, Transformer, and Adaptive Transformer frameworks. Results of the comparison have been provided in table 1.

Figure 2 Comparative Performance of Different Architectures Across Evaluation Metrics



The designed framework yielded accuracy, precision, recall, and F1-Score values of 93.1%, 89.8%, 88.2%, and 89.0%, respectively, with ROC-AUC and PR-AUC metrics being 0.962 and 0.887, respectively, as depicted in figure 2. In particular, the performance improvement was more noticeable in terms of recall and PR-AUC that are suitable for predicting behavioral risk in case of imbalanced datasets due to the fact that they consider how well the model can detect positive cases without relying on the majority class only.

The standard Transformer produced an F1-Score of 82.7%, but by adding adaptive temporal representation, it resulted in an improvement in F1-Score to 85.6%. With the inclusion of the full-fledged neuro-symbolic model, its F1-Score rose to 89.0%.

Calibration Analysis

However, discrimination is not enough when the results of models are probability estimates. Hence, the calibration was assessed based on the Brier score and the expected calibration error (ECE).

Table 2 Calibration Performance		
Model	Brier Score	ECE
Logistic Regression	0.143	0.061
Random Forest	0.129	0.054
XGBoost	0.118	0.047
LSTM	0.112	0.043
Standard Transformer	0.104	0.039
Adaptive Transformer	0.096	0.034
Proposed Framework	0.082	0.027

The proposed model had the lowest Brier score and ECE among all tested. With the Brier score of 0.082, it has better calibration as compared to all other models listed in table 2. Calibration error is significant since the tool was designed to generate risk estimates for professional assessment and not just assign risk classes.

Ablation Study

Ablation experiments were performed to measure the contributions of adaptive temporal learning and symbolic reasoning, which are described in table 3.

Table 3 Ablation Analysis				
Configuration	Accuracy (%)	F1-score (%)	ROC-AUC	
Transformer only	89.2	82.7	0.926	
Transformer + adaptive mechanism	91.0	85.6	0.944	
Transformer + symbolic reasoning	91.8	86.9	0.951	
Adaptive Transformer + symbolic reasoning	93.1	89.0	0.962	

Elimination of the adaptive module lowered the F1 score from 89.0% to 86.9%, while elimination of the symbolic reasoning component, keeping only the adaptive Transformer, lowered the F1 score to 85.6%. The combined model thus improved the F1 score by 3.4% compared to the adaptive Transformer and by 6.3% compared to the regular Transformer.

This shows that the improvement cannot be explained solely by the Transformer model. The adaptive processing of time is useful for representing temporal changes, while the symbolic reasoning gives an additional structured representation of the domain relations.

Temporal Risk Prediction

It is clear from the longitudinal analysis that the risk estimation varied with the order of the observation and not constant throughout the entire one-year period. The average prediction of risk for the entire group increased from 0.21 for the first observation period to 0.34 for the periods with more recent behavioral or contextual changes followed by reduction of the risk once interventions have been performed.

The temporal representation proved to be very sensitive to the most recent observation while taking into account information from past periods. This is indicative of the ability of the adaptive gating approach to help the model take into consideration historical and up-to-date data.

Notably, these probability trends cannot be seen as deterministic predictions of crime. The probabilities are model-based predictions based on the available data to the model.

Explainability Analysis

The global features importance analysis yielded the most relevant feature groups to be recent behavioral change, past behavioral occurrences, contextual instability, communication problems, and changes in support needs. The temporal attribution revealed that recent information was more significant when it depicted a substantial change compared to the prior pattern of the individual. Table 4 illustrates representative global feature importance values.

Table 4		Representative Global Feature Contributions	
Feature		Relative contribution	
Recent behavioural change		0.182	
Previous behavioural incidents		0.161	
Contextual instability		0.137	
Communication-related factors		0.119	
Change in support requirement		0.103	
Environmental stressors		0.091	
Previous intervention response		0.078	
Care-plan changes		0.067	

In this way, the mechanism of explainability offered explanations on both feature and temporal levels. The symbolic part of the mechanism offered another level of explanation that identified the rules of the domain used for the particular prediction. For instance, the activation of the rule on increased preventive monitoring can be triggered by both a recent behavioral change and contextual instability, while the better reaction of the support system can decrease the estimate of risk.

Explanations offer the context for professional examination but do not prove the causality of the factors.

Subgroup Consistency

Subgroup evaluation (Table 5) was conducted to assess demographic stability across distinct cohorts: Group A (individuals requiring high continuous residential support, $n = 135$), Group B (individuals in community-based periodic support, $n = 165$), Age Subgroup 1 (ages 18–34, $n = 160$), and Age Subgroup 2 (ages 35+, $n = 140$). Performance metrics reflect evaluation across individual test splits with 95% confidence intervals.

Table 5	Subgroup Performance			
Group	Accuracy (%)	Recall (%)	F1-score (%)	ROC-AUC
Group A	93.5	88.9	89.6	0.965
Group B	92.7	87.5	88.4	0.959
Age subgroup 1	93.3	88.4	89.2	0.963
Age subgroup 2	92.8	87.8	88.7	0.960

Disparities between the values of ROC-AUC and F1-score in the investigated subgroups were not particularly large. Nonetheless, it is important to take into account the size of each sample and the confidence interval in the context of their interpretation as proof of fairness.

Intervention Planning Results

The intervention module mapped explanations of the model to possible preventive and supportive actions. This mapping was done according to the identified contextual or behavioural factors as well as the symbolic knowledge rules.

As a result, a workflow was generated which gave four main outputs: probability of risk occurrence, significant time-related factors, activated symbolic rules, and possible interventions. Possible interventions involved supportive monitoring, communication support, environment modification, behavioural support, and care plan review. These interventions were not prescribed by the system but rather offered as alternatives for professional evaluation.

This approach converts the model from a pure prediction system into a decision support tool. In addition, it makes it less likely for the risk numerical score to be used separately from the underlying context of the prediction.

Discussion

The results obtained from the experiment highlight the feasibility of using adaptive temporal modeling and neuro-symbolic reasoning for predicting behavioral risk longitudinally in ID care. The performance of the proposed approach was measured in terms of accuracy (93.1%), F1-score (89.0%), ROC-AUC (0.962), PR-AUC (0.887), while Brier score (0.082) and ECE (0.027) were used as measures of calibration in the simulation experiment. F1-score showed an improvement from 82.7% (standard Transformer) to 85.6% (adaptive temporal modeling) and 89.0% (with symbolic reasoning). Recent behavioral change, previous incidents, instability of context, communication issues, and changes in supports were the significant factors for prediction.

The study's implications lie in designing decision-making aids that would link risk estimates from a longitudinal perspective with the context and potential preventive measures. Nonetheless, the study results are based on synthesized data and thus do not validate the framework's efficacy and generalizability. Moreover, the performance of symbolic reasoning depends on the prior defined knowledge base, and the synthetic data does not reflect reality. Recommendations for interventions should be validated by professionals.

There are multiple limitations to the study that need to be taken into account while analyzing the results. First of all, the study relied on the simulated longitudinal data set consisting of 300 individuals and 3,600 observations, thus providing an opportunity for controlled evaluation, yet not accounting for variability, context, documentation, and supporting conditions that apply in ID care in real-life settings. As a result, the conclusions can be viewed as the proof of concept and not the proof of clinical effectiveness of the approach. Furthermore, the binary behavioral risk prediction might be impacted by the differences in documentation and interpretation of the professional. In addition, the framework relies on the symbolic knowledge base that is predefined, and any biases, omissions, and errors could potentially impact the accuracy of predictions. Future research should focus on verifying the framework based on large and diverse longitudinal data sets obtained from real life under ethical governance and performing external and prospective validation with clinicians and caregivers.

Conclusion

The current study proposed an adaptive Transformer-based neuro-symbolic explainable learning framework for longitudinal offender risk prediction and intervention plan development in the context of ID support. The proposed framework combines longitudinal behavioral and contextual data, adaptive temporal modeling, symbolic knowledge, explainable prediction, and intervention-oriented reasoning in a human-in-the-loop decision

support process. According to the experimental results, the proposed framework reached 93.1% accuracy, 89.0% F1-score, 0.962 ROC-AUC, and 0.887 PR-AUC, while the standard Transformer obtained 89.2% accuracy, 82.7% F1-score, and 0.926 ROC-AUC. In addition, the ablation study revealed that adaptive temporal modeling and symbolic reasoning provided an incremental improvement to the predictive ability, with the full framework achieving 6.3 percentage points higher F1-score than the standard Transformer. The calibration study revealed lower Brier score (0.082) and expected calibration error (0.027). The analysis on explainability revealed that recent behavioural change, previous behavioural incidents, context instability, communication aspects, and shifts in support needs were key factors contributing to the models' predictions. These predictions allowed for establishing relationships between risk estimations and potential prevention supports strategies instead of making deterministic decisions on the model's output. Results for subgroups demonstrated generally similar performance in the considered groups; however, further validation is necessary to make conclusions about the framework's generalizability and fairness. Future studies need to conduct an external validation of the framework on independent large longitudinal datasets from various ID-care institutions. The prospective evaluation of the model should focus on such aspects as temporal generalization, calibration drift, subgroup fairness, and robustness to changing conditions of care. It would also be interesting to consider clinicians' and caregivers feedbacks, increase the level of formalization of domain rules, test different approaches to model's explainability, and investigate if intervention suggestions affect the quality-of-care outcomes.

References

1. Alamri, Faisal. "Meta-Adaptive Neuro-Symbolic Causal Graph Transformer for Explainable Brain Tumor Segmentation and Progression Analysis." *Journal of Intelligent Decision Making and Information Science* 3, no. 3s (2026): 1581-1611. <https://doi.org/10.59543/jidmis.v3.957>
2. Wang, L., & Wang, J. (2026). EduRiskNet: Interpretable Early Academic Risk Prediction via State-Aware Trajectory Modeling. *International Journal of Data Warehousing and Mining (IJDWM)*, 22(1), 1-23. <https://doi.org/10.4018/IJDWM.419509>
3. Bettahi, Abdelkarim, Hamid Harroud, and Fatima-Zahra Belouadha. "Early Student Risk Detection Using CR-NODE: A Completion-Focused Temporal Approach with Explainable AI." *Algorithms* 18, no. 12 (2025): 781. <https://doi.org/10.3390/a18120781>
4. Bichi, Abdulkadir Shehu, and Jyoti Shekhawat. "HQIAT-ML: Hybrid Quantum-Inspired Adaptive Transformer with Meta-Learning for Student Dropout Prediction Risk Explanation." In *International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, pp. 759-774. Atlantis Press, 2026. https://doi.org/10.2991/978-94-6239-713-2_56
5. Singh, Harsh, Tanvi Banerjee, and Tanvi Desai. "Neuro-Symbolic AI Framework for Explainable and Reliable Clinical Decision Support." *International Journal of Unified Research & Development (IJURD)* 2, no. 2 (2026). <https://doi.org/10.5281/ijurd.v2i2.30>
6. Jena, Monalisa, Noman Khan, Mi Young Lee, and Seungmin Rho. "Neuro-Symbolic Graph Learning for Causal Inference and Continual Learning in Mental-Health Risk Assessment." *Computer Modeling in Engineering & Sciences* 146, no. 1 (2026): 44. <https://doi.org/10.32604/cmescs.2025.075119>
7. Prajwalasimha, S. N., Nilesh Shelke, Dilip Kumar Jang Bahadur Saini, Amit Pimpalkar, Sharanabasappa Tadkal, and Rajeshwar Balla. "Hybrid Transformer-CNN Neuro-Symbolic Explainable AI for Cyber Threat Intelligence: Advancing Transparency and Adversarial Robustness." In *2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoCI)*, pp. 492-499. IEEE, 2025. <https://doi.org/10.1109/ICoCI65217.2025.11254796>
8. Silva, Izabella, João Paiva, Mirko Perkusich, Danylo Albuquerque, Emanuel Dantas, Kyller Gorgônio, and Angelo Perkusich. "Evaluating the Quality of User Stories: An Extended Comparative Study of Multiple LLMs and Rule-Based Tools." In *International Conference on Agile Software Development*, pp. 155-174. Cham: Springer Nature Switzerland, 2026. https://doi.org/10.1007/978-3-032-22375-3_10
9. Jammal, Fares, Mohamed Dahab, and Areej Y. Bayahya. "Neuro-bridge-X: A Neuro-symbolic vision transformer with meta-XAI for interpretable leukemia diagnosis from peripheral blood smears." *Diagnostics* 15, no. 16 (2025): 2040. <https://doi.org/10.3390/diagnostics15162040>

10. Alzahrani, Ahmed. "Cognitive ZTNA: A Neuro-Symbolic AI Approach for Adaptive and Explainable Zero Trust Access Control." *Mathematics* 14, no. 7 (2026): 1211. <https://doi.org/10.3390/math14071211>
11. Okuwobi, Idowu Paul, Jingyuan Liu, Olayinka Susan Raji, and Olusola Funsho Abiodun. "Adaptive Neuro-Symbolic framework with dynamic contextual reasoning: A novel framework for semantic understanding." *Mathematical Biosciences and Engineering: MBE* 22, no. 12 (2025): 3028-3059. <https://doi.org/10.3934/mbe.2025112>
12. Ibrahim, Zahraa Sameer, Haedar Ahmed Mukhef, and Hayder Hasan Ali. "Memory-Efficient Probabilistic Neuro-Symbolic Integration for Explainable Natural Language Inference Using Transformer-Based Foundation Models." *Al-Mustansiriyah Journal of Science* 37, no. 2 (2026): 29-42. <https://doi.org/10.23851/mjs.v37i2.1871>
13. Jahani-Nasab, Mahyar, Hamid Reza Tofighi, Mor Vered, and Buser Say. "PING: A Physics-Informed Neuro-Symbolic Generator for Continuous-Time Planning." In *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 36, no. 1, pp. 113-122. 2026. <https://doi.org/10.1609/icaps.v36i1.42820>
14. Fitas, Ricardo. "Neuro-Symbolic AI for Personalized and Equitable Virtual Tutoring in Education." In *Virtual Tutors and AI-Powered Instructional Tools in K-12 Settings*, pp. 1-64. IGI Global Scientific Publishing, 2026. <https://doi.org/10.4018/979-8-3373-2637-5.ch001>
15. Nicolau, Marcio, Claudio R. Jung, and Luis C. Lamb. "Neurosymbolic computer vision: a survey and perspective." *Neural Computing and Applications* 38, no. 14 (2026): 574. <https://doi.org/10.1007/s00521-026-12063-6>
16. Sendi, Asaad, Foyzul Hoque, and Md Tamjidul Hoque. "Decoding Stock Trends: A Comparative Study of GRU LSTM and Transformer Models in Tech Sector Prediction." *Trans. Eng. Comput. Sci.* 13, no. 03 (2025): 46-60. <https://doi.org/10.14738/tmlai.1303.18843>
17. Bowring, Darren L., Vasiliki Totsika, Richard P. Hastings, Sandy Toogood, and Gemma M. Griffith. "Challenging behaviours in adults with an intellectual disability: A total population study and exploration of risk indices." *British Journal of Clinical Psychology* 56, no. 1 (2017): 16-32. <https://doi.org/10.1111/bjc.12118>
18. Marrus, Natasha, and Lacey Hall. "Intellectual disability and language disorder." *Child and adolescent psychiatric clinics of North America* 26, no. 3 (2017): 539. <https://doi.org/10.1016/j.chc.2017.03.001>
19. Rajendren, D. J., M. Sheetal Kumar, D. Rajesh, Z. Balassem, I. Juraeva, and K. K. S. Ugli. "Named Entity Linking for Cross-Document Information Extraction." *Indian Journal of Information Sources and Services* 15, no. 3 (2025): 413–422. <https://doi.org/10.51983/ijiss-2025.IJISS.15.3.46>
20. Patel, Dilip R., Maria Demma Cabral, Arlene Ho, and Joav Merrick. "A clinical primer on intellectual disability." *Translational pediatrics* 9, no. Suppl 1 (2020): S23. <https://doi.org/10.21037/tp.2020.02.02>
21. Mirdjonovna, Khamroeva Shahlo, Gulyamova Shakhnoza Kakhramonovna, Eshmuminov Askar Allamuradovich, Akhmedova Dildora Dilshodovna, Rakhimova Sulurkhan Djumaniyozovna, Kholova Muyassar Abdulhakimovna, and Karimov Rustam Abdurasulovich. "80/20 Principle in Phonetics, Lexis and Word Formation." In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pp. 433-438. IEEE, 2025. <https://doi.org/10.1109/UBMK67458.2025.11207076>