International Journal of Medical Toxicology & Forensic Medicine
2026, Volume 16, E53727

Research Paper

AI-Driven Predictive Analytics for Early Detection of Mental Health Disorders in Criminal Justice Systems

Sabarinathan Gurunathan[1], Aavleen Bakshi[2], Garaga Srilakshmi[3], Irissappane Dhanusu Soubache[4], R Rajagopal[5*]

1. *Department of Mathematics, SRM Bharathidasan College of Engineering and Technology, Pudukkottai, India.*
2. *Department of Psychology, Christ (Deemed to be University), Delhi NCR Campus, Ghaziabad, India.*
3. *Department of Electronics and Communication Engineering, Aditya University, Surampalem, Andhra Pradesh, India.*
4. *Eudoxia Research University, New Castle, USA.*
5. *Department of Computer Science and Engineering, Sree Rama Engineering College, Tirupati, India.*

Citation Gurunathan S, Bakshi A, Srilakshmi G, Soubache ID, Rajagopal R. AI-Driven Predictive Analytics for Early Detection of Mental Health Disorders in Criminal Justice Systems. International Journal of Medical Toxicology and Forensic Medicine. 2026; 16:E53727.

https://doi.org/10.22037/ijmtfm.v16.53727

Article info:

Received: 10 October, 2026

Accepted: 05 November, 2026

Published: 09 November, 2026

Keywords:

Machine Learning, Explainable AI, Quick Medical Reference (QMR), Mental Health Disorder (MHD), Criminal Justice System (CJS)

ABSTRACT

Background: Mental health disorders among individuals involved in criminal justice systems require timely and reliable assessment. Automated predictive approaches may assist mental health screening by integrating behavioural, clinical, and textual information. This study proposes an explainable hybrid framework combining a Bidirectional Encoder Representations from Transformers (BERT)-based language model with a Quick Medical Reference (QMR)-style Bayesian network for mental health disorder prediction.

Methods: The proposed framework incorporates data preprocessing, class balancing, dimensionality reduction, and hybrid probabilistic reasoning. Synthetic Minority Over-sampling Technique (SMOTE) is applied to address class imbalance, while Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are employed for dimensionality reduction. The BERT-based model generates initial disorder probabilities from symptom descriptions, which are subsequently refined using a QMR-style Bayesian network through iterative symptom-based evidence updating. The proposed framework is evaluated using the Survey of Inmates in Local Jails (SILJ) 2024–2025 dataset and compared with Logistic Regression, XGBoost, and Explainable Boosting Machine.

Results: The proposed BERT-QMR-style Bayesian framework achieved an accuracy of 95.08%, precision of 93.41%, recall of 91.62%, F1-score of 92.51%, and AUC-ROC of 98.34%. The model also achieved an RMSE of 0.074 with a computational time of 64.2 seconds. Compared with the performance before preprocessing, the proposed framework improved accuracy from 86.76% to 95.08% and AUC-ROC from 89.63% to 98.34%, demonstrating the potential contribution of the integrated preprocessing and hybrid prediction strategy.

Conclusion: The proposed hybrid framework demonstrates the potential of combining contextual language representation with interpretable Bayesian reasoning for mental health assessment in criminal justice settings. The integration of BERT, SMOTE, PCA, ICA, and QMR-style Bayesian inference provides a promising approach to predictive mental health screening. Further validation using independent datasets, clinically validated labels, and external evaluation is necessary before practical deployment.

*** Corresponding Author:**
Rajagopal R, PhD
Department of Computer Science and Engineering, Sree Rama Engineering College, Tirupati, India.
E-mail: *rajagopalrmail@gmail.com*

Copyright © 2026 The Author(s);
This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0: https://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Obesity is Mental health diagnoses are frequently regarded by courts as either aggravating or mitigating considerations during sentencing proceedings. States' varying laws and the way they are applied frequently result in disparities in the frequency and duration of incarceration of people with mental illnesses in the country. Sentencing legislation and guidelines in places with more conservative political ideologies have specifically been proposed to seriously disfavor those with psychiatric disorders [1]. Many individuals involved with the criminal justice system struggle with both mental illness and substance use issues. However, there is a lack of data about the likelihood of arrest or incarceration for those with CODs. This is due to the fact that the majority of the existing literature exclusively addresses individuals with severe mental illness or substance use disorders [2]. As a result, they looked at the clinical and sociodemographic factors that are most associated with CJSI in COD patients, and then compared the rates of CJSI among patients with severe mental illness and those with drug use disorder to determine which risk factor is more significant.

Mental health Disorders in the criminal justice system can greatly benefit from AI, as seen in Figure 1 [3]. Increased accuracy and efficiency, objective analysis, early issue discovery, and better resource allocation are the four key advantages that are highlighted.

Recent advances in artificial intelligence and machine learning have enabled the development of automated approaches for mental health prediction [4]. Machine learning models have been applied to clinical records, behavioral information, physiological signals, and textual descriptions to identify patterns associated with depression, anxiety, stress, and other mental health conditions [5]Deep learning and language models can further extract semantic information from natural-language descriptions of symptoms [6]**.** However, conventional predictive models may provide limited interpretability, which is particularly important when AI systems are applied to sensitive healthcare and criminal justice settings [3].

Bayesian networks provide an interpretable probabilistic framework for representing relationships between disorders and associated symptoms. They can incorporate prior clinical knowledge and update diagnostic probabilities when new evidence becomes available. However, Bayesian models may depend on predefined symptom representations and may not directly process free-form natural-language descriptions [7].

To address these limitations, this study proposes a hybrid BERT-QMR-style Bayesian framework that combines language-based symptom representation with probabilistic Bayesian reasoning. The primary contribution of this study is the integration of contextual language representation with probabilistic symptom–disorder reasoning in a unified predictive framework.

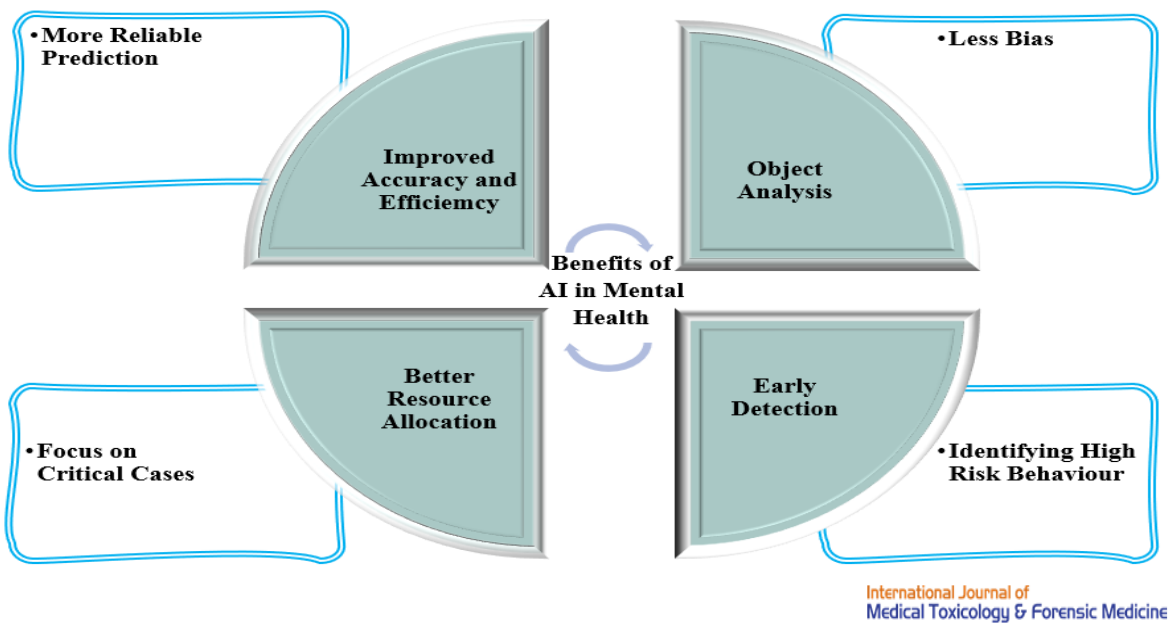


Figure 1. Schematic Representation of AI Representation in Mental Health Disorders in Criminal Justice System.

Unlike approaches that rely exclusively on conventional classification or predefined symptom indicators, the proposed BERT-QMR-style architecture uses BERT to derive initial disorder probabilities from textual symptom descriptions and a QMR-inspired Bayesian network to update these probabilities using additional symptom evidence. The framework also investigates the influence of class balancing and dimensionality reduction on predictive performance. These contributions are intended to support transparent preliminary mental health screening rather than replace professional clinical assessment.

Machine Learning for Mental Health Prediction

Machine learning techniques have been increasingly investigated for the prediction and assessment of mental health conditions [4]. Conventional approaches including Logistic Regression, Support Vector Machines, Random Forest, and boosting algorithms have been applied to clinical, behavioral, and textual data [5]. These models can identify statistical relationships between observed characteristics and mental health outcomes [8]. However, their predictive performance may depend strongly on feature quality, class distribution, and the availability of representative training data [4, 9].

Deep Learning and Language-Based Mental Health Analysis

Deep learning models have enabled automated representation learning from complex mental health data [10]. CNN, RNN, LSTM, and transformer-based architectures have been used to analyze textual and behavioral information associated with depression, anxiety, stress, and other disorders [4]. Transformer-based language models such as BERT can capture contextual relationships between words and provide richer representations of natural-language symptom descriptions [6]. Nevertheless, deep learning predictions can be difficult to interpret, particularly when used in sensitive healthcare and criminal justice applications [3].

Bayesian Models for Explainable Mental Health Assessment

Bayesian networks provide a probabilistic mechanism for representing relationships between diseases, symptoms, and clinical evidence [11]. Their graphical structure can provide greater interpretability than purely black-box predictive models [12]. QMR-style Bayesian reasoning further allows disease–symptom relationships to be represented through conditional probabilities [13]. However, traditional Bayesian approaches generally require structured symptom information and may not directly process free-form symptom descriptions [14].

Research Gap

Existing studies demonstrate the effectiveness of machine learning, deep learning, and Bayesian reasoning for mental health assessment [4]. However, most approaches either focus on predictive modeling without explicit probabilistic reasoning or use Bayesian models without directly integrating contextual language representations. In addition, behavioral and temporal characteristics are not consistently incorporated with textual symptom information. Therefore, there is a need for an integrated framework that combines contextual language representation and interpretable probabilistic reasoning [15, 16]. The proposed BERT-QMR-style Bayesian framework addresses this gap by using BERT to generate initial disorder probabilities and a Bayesian network to refine these predictions using symptom-level evidence.

Materials and Methods

There is widespread concern about the increasing number of individuals with mental health problems who are involved with criminal justice systems around the world. There is evidence that these individuals are disproportionately represented in prison populations and the criminal justice system as a whole, in roles such as victims and offenders. Having a significant number of patients with MHD housed in the CJS has enormous personal, societal, economic, and human implications. Now more than ever before, it is crucial to conceptualize the intricate web of social processes, systemic responses, and individual impairment experiences that makes people with MHD more susceptible to encountering the CJS and their varied journeys within, outside, and frequently circling back to these systems.

Figure 2 shows a conceptual framework for using AI in Criminal Mental Health Risk Assessment. It shows a systematic approach from data preprocessing to predictive analysis. The first step is to assess the risk. Then, preprocessing methods like SMOTE, PCA, and ICA are used to address data imbalance and reduce dimensionality. After that, an AI-powered predictive algorithm uses models like XGBoost, Explainable Boosting Machine, BERT-QMR Style Bayesian Network, and Logistic Regression to process the extracted characteristics. Finally, the model's performance is tested against important metrics like accuracy, precision, recall, F1-score, AUC, RMSE, and computational time. This makes sure that the predictions about criminal mental health risks are reliable and easy to understand.

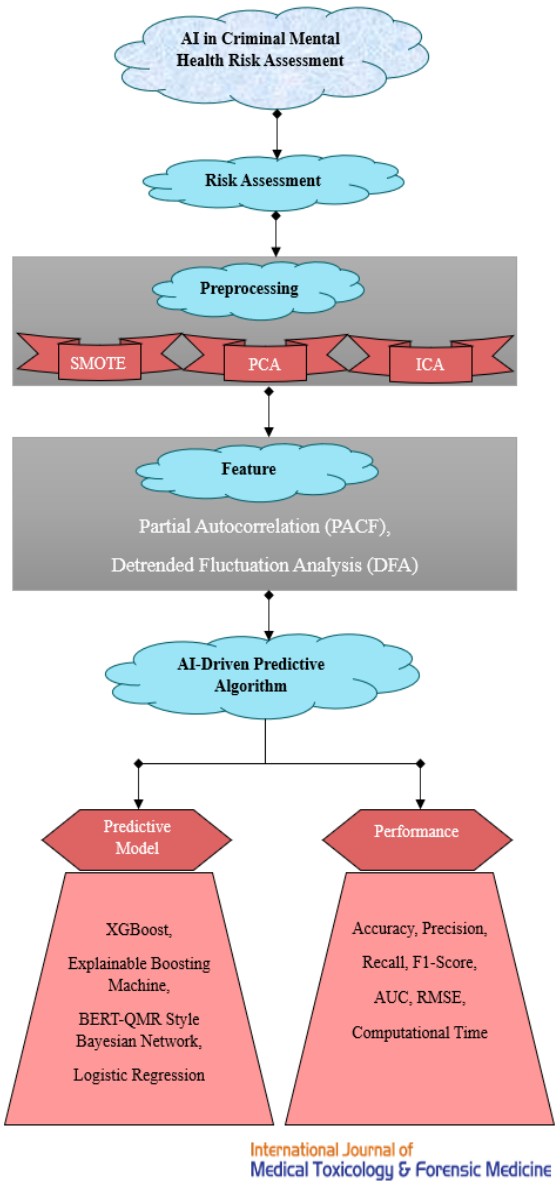


Figure 2. Schematic representation of AI representation in mental health disorders in criminal justice system.

Preprocessing

To aid in the detection of stress and associated MDs, this section reviews key preprocessing strategies that have produced substantial results.

Class Imbalance Using Synthetic Minority Oversampling Technique

The number of samples for the stress or MHD class is typically much fewer than the non-stress or non-MHD class when it comes to stress and its related MHD detection. Predictions are skewed since there aren't enough samples from each class. It is feasible to oversample the under-represented group in order to rectify data bias[5]. To reduce the influence of class imbalance, SMOTE is applied to the training data. SMOTE generates synthetic minority-class observations by interpolating between existing minority samples.

PCA Method

PCA is used to reduce the dimensionality of these types of data sets while keeping as much information as possible and making them easier to understand by using principal component analysis (n=3, 3%). The transformed components are subsequently used as compact representations for predictive modeling.

ICA

A computational and statistical technique for discovering hidden components underlying random variables, observations, or signals is independent component analysis (ICA; n=4, 4%). Removing artifacts from multichannel data that contains stationary signal noises is the primary utility of this technology. PCA optimizes the data's covariance matrix, which represents second-order statistics, whereas ICA optimizes higher-order statistics like kurtosis. The resulting representations are used to capture underlying independent patterns that may contribute to mental health prediction.

Model Training

Modeling the Bayesian Network

Our goal in building this Bayesian network was to simulate the diagnostic process by capturing the links between symptoms and conditions. By confirming or denying the detected symptoms, this network is meant to refine the diagnosis that was derived from the language model. The nodes in the network are categorized as either symptoms (S) or mental health (D). The nodes for symptoms indicate whether symptoms are present or not, whereas the nodes for mental health disorders indicate if mental health disorders are present or not [11]. A Conditional Probability Table (CPT) is established at each edge connecting the nodes, which represents the conditional dependencies between symptoms.

The Quick Medical Reference (QMR) network, a Bayesian network tailored for medical diagnostics, is the foundation of our network. It uses a two-layer structure to depict the linkages between MHD and symptoms. On the top layer are nodes representing MHD, and on the bottom layer are nodes representing symptoms or exam results. From MH to findings, there are arcs.

Most of the time, the QMR network represents a joint probability distribution for all MH and symptoms in Equation (2).

$$N(Q,B) = N(Q \mid B)N(B) = [\prod_(k = 1)\^p \llbracket N(q_k \mid b) \rrbracket][\prod_(l = 1)\^o N(b_l)] \quad (2)$$

$B = \{b_1, b_2, \ldots, b_o\}$ and $Q = \{q_1, q_2, \ldots, q_p\}$ are binary variables that represent the existence or absence of specific symptoms or mental health issues, respectively.

Keep in mind that m conditions $\{b_1, \ldots, b_o\}$ can cause a symptom q. $2\^m$ parameters are needed to construct this conditional probability distribution. Since an expert would have to think about every possible combination of parent variable activation for each symptom, this presents a huge hurdle that makes experts' jobs more difficult when defining probability.

A more natural course of action for professionals would be to calculate the sensitivity and specificity scores for each sign and disease. According to the equation (3), sensitivity is the likelihood that a symptom will manifest when a condition b_k is present, given a symptom q_k and a condition b_l:

$$Qg_kl \coloneqq N(q_k = 1 \mid b_l = 1) \quad (3)$$

If the condition b_l is not present in equation (4), then the likelihood that symptom b_k does not manifest is known as specificity.

$$Qn_kl \coloneqq N(q_k = 0 \mid b_l = 0) \quad (4)$$

The Noisy-OR model was proposed to represent these probabilities. In this approach, the likelihood of a symptom being present is the sum of the probabilities of each circumstance acting alone to generate the symptom. According to mathematical assumptions, there are o associations between the parent conditions $\llbracket \{b\} \rrbracket _k, \ldots, b_o\}$ and the symptom q, which are combined in an OR gate. Each parent-symptom link has a $1 - n_k$ chance of failing in order to incorporate stochasticity into the model. The likelihood of the symptom being "on" is provided by under these assumptions in equation (5).

$$N(q_k = 1 \mid b) = 1 - \prod_(l = 1)\^o (1 - n_kl) \quad (5)$$

The probability that the condition b_l could independently trigger the symptom q_k, if it were present, is represented by n_kl here. In order to build the network, the Noisy-OR model is used to define the links between symptoms and illnesses. What follows is an explanation of how the MH is fine-tuned using a combination of the language model's predictions and those from the Bayesian network.

The QMR-style Bayesian component represents mental health categories and their associated symptoms as probabilistic variables. The network structure defines the relationships between disorder nodes and symptom nodes, while conditional probability parameters quantify the strength of these relationships. When additional symptom evidence becomes available, posterior probabilities are updated according to the specified Bayesian inference procedure. If a Noisy-OR formulation is used, its assumptions, leak probabilities, link parameters, and method of parameter estimation must be stated.

A Hybrid Approach for Mental Health Refinement

BERT is employed to capture contextual relationships within natural-language symptom descriptions. The text is tokenized using the selected pretrained BERT tokenizer and converted into token identifiers and attention masks. The resulting contextual representation is passed through a classification layer to obtain initial probabilities for the predefined target categories. Structured demographic, behavioural, and clinical variables require a separately specified encoding procedure if they are included in the model. The study should identify the exact BERT checkpoint, maximum sequence length, training strategy, optimizer, learning rate, batch size, and model-selection criteria to support reproducibility.

Let's now lay forth the mathematical foundations of our symptom detection algorithm. Imagine a network E that includes a set of MHD $B = \{b_1, \ldots, b_o\}$ and symptoms $Q = \{q_q, \ldots, q_p\}$. In this network, each disease b_k is linked to a subset of symptoms q_k, and the Edges between symptoms q_k and their associated diseases b_l are defined by a Conditional Probability Table (CPT) that follows $N(q_k \mid b_l)$ the Noisy-OR assumptions.

A BERT model calculates the a priori probabilities of the MH vector cap N open paren cap B vertical bar when a user submits a list of words cap R equals open paren u sub 1, dot dot dot ,u sub r close paren to describemits a list of words $R = (u_1, \ldots, u_r)$ to desribe symptoms. Te function $h_\theta(R)$, where θ stands for the model weights, reflects the probability. Equation (6) is used to formalise the BERT model output.

$$h_\theta\ (R) = softmax\ (U\hat{}R \cdot BERT_enc\ (R)) \quad (6)$$

Considering$U \in T\hat{}(o \times F)$where H is the size of the output vector produced by the BERT model.

Contrary to our deep learning model's category prediction capabilities, the Bayesian network handles individual conditions, as mentioned before. In order to solve this, they find the I most likely categories based on what the neural network predicted. Then, using the BERT model's output probabilities for each category, they build a graph with the top i conditions linked to these categories.

They begin by selecting the top i illnesses from the MH probability vector $N(B \mid R)$, as previously stated, and then use the a priori values and the network E to calculate the symptom probability in the second step of the process in equation (7).

$$B_i \leftarrow top_i\ (N(B \mid R)) \quad (7)$$

By doing so, they may create a subnetwork E_i from E that exclusively contains the MH in B_i and the symptoms that are connected with them, $Q_i = \cup_(b_k \in B_i) q_(b_k)$.

As seen in equation (8), this subnetwork represents the joint distribution.

$$N(Q,B) = N(Q \mid B)N(B) = \prod_(q_k \in Q_i) [\![N(q_k \mid B)]\!] \prod_(b_l \in B_i) [\![N(b_l \mid R)]\!] \quad (8)$$

Using Equation (9), we can calculate the likelihood of each symptom.

$$N(q_k \mid R) = \sum_(b \in B_i) [\![N(q_k \mid b)N(b \mid R)]\!] \quad (9)$$

Equation (10) calculates the odds of each mental health state.

$$N(g_l \mid Q) = \sum_(b \in B_i) [\![N(Q \mid b_k)N(b_k)]\!] \quad (10)$$

where P is the normalization factor that guarantees the total of the probabilities is 1.

To begin collecting fresh evidence, we ask the user to affirm or refute the existence of the q most prominent symptoms listed in equation (11):

$$Q_s \leftarrow top_s\ (N(Q \mid R)) \quad (11)$$

For each stage of the algorithm, s denotes the number of symptoms that the user confirms or rejects, and Q_s stands for the symptoms with the highest updated probabilities. After that, at every stage, the user presents evidence that appears to be$g = \{Q_1 = 1, Q_2 = 1, \ldots, Q_s = 0\}$.

The MHD probability vector is updated after adding the new evidence e to the current evidence vector $G = G + e$.

The procedure then repeats for r *iterations, updating the probabilities and obtaining fresh evidence each time, when new probabilities*

Algorithm 1. Algorithm for the proposed BERT-QMR style Bayesian network.

	Algorithm 1: Algorithm For BERT-QMR
	Input: Signs and symptoms word vector $R \leftarrow (u_1, \ldots, u_r)$
	Output: Predicted mental health disorder B_n
1.	**Procedure** BERT-QMR:
2.	Get the BERT model initiated by adding weights θ.
3.	Let $B \leftarrow \{b_1, .., b_o\}$ be the set of mental health.
4.	Let $Q \leftarrow \{q_1, \ldots, q_p\}$ be the set of Symptoms.
5.	The mental health disorder-symptoms network is denoted as E.
6.	Compute $n(B\|R)$ using BERT:$h_\theta(R) \leftarrow softmax\left(U^R BERT_{enc}(R)\right)$.
7.	Select top i mental health: $B_i \leftarrow top_i\left(N(B\|R)\right)$
8.	Construct subnetwork E_i from E using B_i
9.	Initialize evidence as $G \leftarrow \{\}$
10.	**For** iteration $k \in \{1 \ldots r\}$ **do**
11.	Compute symptom probabilities as $N(q_k\|R) \leftarrow \sum_b N(q_k\|b)N(b\|R)$
12.	Select r symptoms using $Q_s \leftarrow top_s(N(Q\|R))$
13.	Ask user to provide evidence g based on Q_s
14.	Update evidence vector $G \leftarrow G + g$
15.	Update $N(b_l\|Q,G)$ for each mental health disorder b_l in E_i.
16.	$N(b_l\|Q) \leftarrow \frac{1}{P}\sum_{b \in B_i \setminus \{b_l\}} N(Q\|b)N(b)$
17.	**End for**
18.	Final prediction: $B_n \leftarrow \arg\max(N(B\|R,G)$

for the symptoms eliminated from the evidence vector $Q_p\hat{}' = Q - G$ are calculated. Equation (12) concludes with the most likely MHD.

$$B_n \leftarrow \arg[\![\max(N(B|R,G)]\!] \quad (12)$$

Algorithm 1 continues the full algorithm.

Results

Persistent emotional and mental health issues can be a lasting consequence of being a victim of violent crime. These symptoms can manifest as real mental health issues in certain people. In other cases, the effects are more nebulous, but victims nevertheless report increased anxiety, anguish, and a worse quality of life. Research on the effects of interactions with the criminal justice system on the psychological well-being of adult victims of crime. The majority of criminal justice organizations lack the resources necessary to help victims cope with their emotional distress. Victims' needs are frequently ignored because court and police personnel are under-trained to identify and handle mental health issues.

Table 1. Summary of the dataset.

Feature Category	Example Attribute	Description/ Role of the Model
Demographics	Age, Gender, Race, Education	Cultural Risk Factors and Socioeconomic.
Criminal History	Types of Offense, Sentence Duration, Prior Arrests.	Situational Predictors and Behavioral.
Health Indicators	Prior Mental Health Diagnosis, Medication Usage, Substance Abuse	Core Diagnostic Correlates.

International Journal of
Medical Toxicology & Forensic Medicine

With an emphasis on demographic, behavioral, criminal, and psychological factors, the SILJ 2024–2025 dataset offers extensive data at the inmate level throughout local jails in the United States. This dataset provides a solid basis for future research into the causes and risk factors of mental health issues among people who have had contact with the legal system (Table 1) [17].

The Figure 3 illustrate the discrimination performance of the evaluated models. The proposed BERT-QMR-style Bayesian framework exhibits the highest reported AUC among the compared approaches. However, the ROC calculation procedure and averaging strategy should be specified for the multiclass prediction setting.

Table 2. Model evaluation.

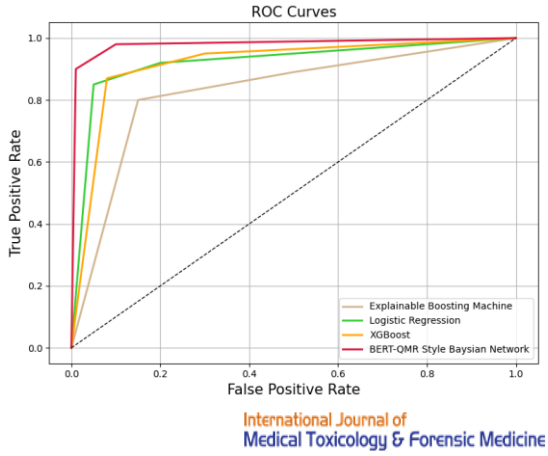
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	90.76	88.12	86.23	90.97
XGBoost	93.42	91.30	89.11	93.88
BERT-QMR Style Bayesian Network	95.08	93.41	91.62	92.51
Explainable Boosting Machine	87.12	85.71	83.37	87.80

International Journal of
Medical Toxicology & Forensic Medicine

Table 3. Model evaluation for before and after preprocessing.

Metrics	Before Preprocessing	After SMOTE+PCA +ICA	Proposed Model (BERT-QMR Style Bayesian Network
Accuracy	86.76	90.37	95.08
Precision	85.30	89.31	93.41
Recall	83.17	87.12	91.62
F1-Score	86.81	90.05	92.51
AUC-ROC	89.63	94.32	98.34
RMSE	0.176	0.142	0.074
Computational Time	47.5	55.7	64.2

International Journal of
Medical Toxicology & Forensic Medicine



International Journal of
Medical Toxicology & Forensic Medicine

Figure 3. Comparison of error metrics.

Table 2 presents the comparative performance of four machine learning models: Logistic Regression, XGBoost, BERT-QMR Style Bayesian Network, and Explainable Boosting Machine in terms of accuracy, precision, recall, and F1-score. The proposed BERT-QMR Style Bayesian Network achieved the highest performance across all four-evaluation metrics, with an accuracy of 95.08%, precision of 93.41%, recall of 91.62%, and F1-score of 92.51%. XGBoost achieved the second-highest performance, followed by Logistic Regression, whereas the Explainable Boosting

Machine obtained comparatively lower results. These findings indicate that the proposed hybrid architecture provides a stronger balance between precision and recall and demonstrates improved classification performance compared with the baseline models.

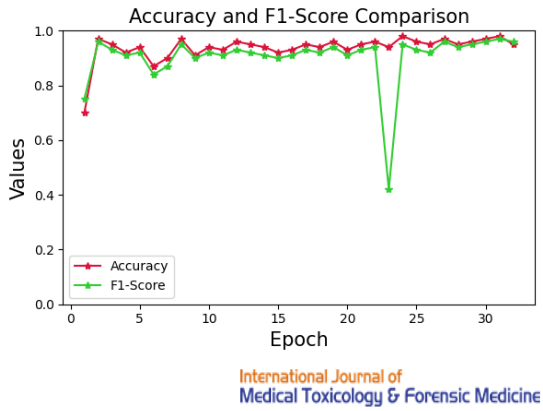


Figure 4. Training accuracy and F1-score across 32 epochs for the BERT-based component.

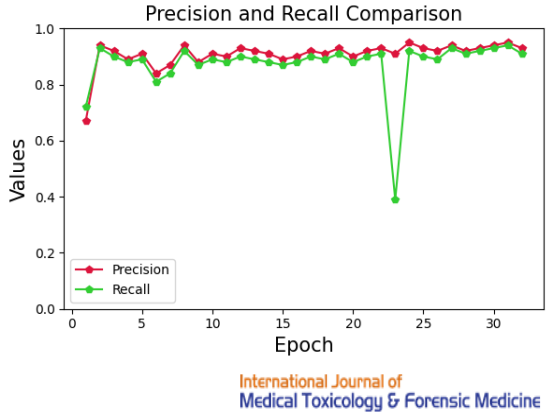


Figure 5. Precision and recall comparison across training epochs.

Figure 4 shows a comparison of the accuracy and F1-score over 32 epochs of model training. Throughout the training phase, both metrics stay high, with accuracy (red line) marginally better than the F1-score (green line). There is a sudden drop in the F1-score around the 23rd epoch, which means that precision and recall are temporarily out of balance. However, it quickly returns to stable values. Overall, the graph shows that the model does well and is reliable over time, with only small changes.

Figure 5 shows a comparison of precision and recall over 32 epochs of model training. Both metrics indicate good and constant performance across the epochs. The red line for precision always has values that are somewhat higher than the green line for recall. Around the 23rd epoch, recall drops noticeably, but it quickly goes back up, showing that the model's sensitivity changes for a short time. The graph shows that the model works well overall, with consistent precision and recall trends across training cycles.

Table 3 shows how the suggested BERT-QMR Style Bayesian Network model gets better at different phases of preprocessing. Before preprocessing, the model did okay, with an accuracy of 86.76% and an AUC-ROC of 89.63. After using SMOTE, PCA, and ICA, the metrics got a lot better, showing that the data was better balanced and the features were better extracted. The last proposed model did quite well, getting an accuracy of 95.08%, a precision of 93.41%, a recall of 91.62%, and an F1-score of 92.51%. It also had the lowest RMSE (0.074) and the highest AUC-ROC (98.34). The computational time did go up a little bit (64.2 seconds), but the overall performance gains show that the model is better at classifying things and more reliable.

The BERT-QMR Style Bayesian Network has the best True Positives and True Negatives while also having the fewest False Positives and False Negatives. This shows that the model works better than others. Linear Regression, on the other hand, is less accurate and has more mistakes. In general, the results show that the suggested BERT-QMR Style Bayesian Network model works better and is more reliable than traditional and boosting-based methods, as shown in Figure 6.

Discussion

The findings of this study indicate that integrating contextual language representation with probabilistic Bayesian reasoning has potential for improving preliminary mental health disorder prediction in criminal justice settings. The proposed BERT-QMR-style Bayesian framework achieved a reported accuracy of 95.08%, outperforming Logistic Regression (90.76%), XGBoost (93.42%), and Explainable Boosting Machine (87.12%). This improvement suggests that combining BERT-based contextual symptom representation with Bayesian evidence updating may provide an advantage over approaches that rely exclusively on conventional classification techniques. The results are consistent with the growing application of artificial intelligence and machine learning in mental health assessment, where complex behavioural and textual patterns can support the identification of potential psychological concerns.

An important contribution of the proposed framework is its integration of language-based prediction with symptom-level probabilistic reasoning. BERT extracts contextual information from natural-language symptom descriptions and generates initial disorder probabilities. The QMR-style Bayesian network

subsequently updates these probabilities using additional symptom evidence. This combination addresses a limitation of conventional language models, which may provide predictions without sufficiently transparent reasoning. Bayesian networks represent relationships between symptoms and disorders through conditional probabilities, allowing the prediction process to be examined and updated when new evidence becomes available. Similar benefits of combining deep learning and Bayesian reasoning have been discussed in previous research on explainable mental health diagnosis .

The preprocessing results also demonstrate the potential contribution of class balancing and dimensionality reduction. As reported in Table 3, accuracy increased from 86.76% before preprocessing to 90.37% after the application of SMOTE, PCA, and ICA, followed by a reported improvement to 95.08% with the proposed hybrid framework. The AUC-ROC increased from 89.63% to 94.32% and subsequently to 98.34%, while RMSE decreased from 0.176 to 0.142 and then to 0.074. These changes indicate that the preprocessing and hybrid modelling stages may contribute to improved discrimination and reduced prediction error. Nevertheless, the individual effects of SMOTE, PCA, and ICA should be evaluated separately through ablation experiments to establish the contribution of each component.

The comparative results further indicate that XGBoost achieved the second-highest reported accuracy, followed by Logistic Regression and Explainable Boosting Machine. This observation suggests that ensemble-based learning can capture useful relationships in structured data, while the proposed hybrid approach may benefit from combining these relationships with contextual symptom information. Previous research has similarly explored transformer-based methods for behavioural health classification using police narratives, demonstrating the relevance of contextual language processing in criminal justice applications. Research on explainable machine learning for mental health prediction has also highlighted the importance of model transparency and the interpretation of behavioural indicators.

The reported AUC-ROC of 98.34% indicates strong discrimination between the target classes in the evaluated dataset. However, high discrimination alone does not establish that predicted probabilities are well calibrated or that the model will perform consistently on data from different criminal justice institutions. Calibration analysis, class-wise sensitivity and specificity, confusion matrices, and external validation are therefore necessary to assess the reliability of the proposed framework. In particular, the performance should be examined across demographic groups to identify possible differences in prediction errors. Fairness is especially important in criminal justice applications because automated predictions may influence decisions involving individuals from diverse social and demographic backgrounds.

The dataset used in this study contains demographic characteristics, criminal history information, and health indicators, as summarized in Table 1. These categories provide a potentially useful foundation for examining relationships between individual circumstances and mental health outcomes. However, the dataset description and target-label construction require additional clarification. The manuscript should explain how the mental health disorder labels were obtained, whether they were clinically validated, how textual symptom descriptions were associated with the structured records, and which variables were supplied to each model. Such details are essential to establish that the model is learning meaningful clinical patterns rather than relying on indirect or potentially biased associations.

The training curves presented in Figures 4 and 5 show generally stable performance across the reported 32 epochs, although a noticeable reduction in F1-score and recall occurs around the 23rd epoch. This temporary variation may indicate fluctuations in the balance between precision and recall during training. The final model-selection procedure should therefore be clearly documented, including the validation strategy, stopping criterion, selected epoch, and whether the reported results were obtained from an independent test set. Further evaluation using stratified cross-validation and an external dataset would help establish the robustness and reproducibility of the proposed framework.

Despite these promising findings, the proposed system should be considered a preliminary screening and decision-support approach rather than an automated diagnostic or criminal justice decision-making tool. Mental health assessment requires professional clinical evaluation, contextual understanding, and consideration of individual circumstances. The use of sensitive demographic, criminal history, and health-related information also raises important concerns regarding privacy, informed governance, fairness, and the potential misuse of predictive outputs. Future work should therefore focus on clinically validated labels, independent multi-institutional datasets, transparent feature attribution, probability calibration, and fairness evaluation. Overall, the proposed BERT-QMR-style Bayesian framework provides a potential direction for integrating contextual language understanding and interpretable probabilistic reasoning, subject to further methodological verification and external validation.

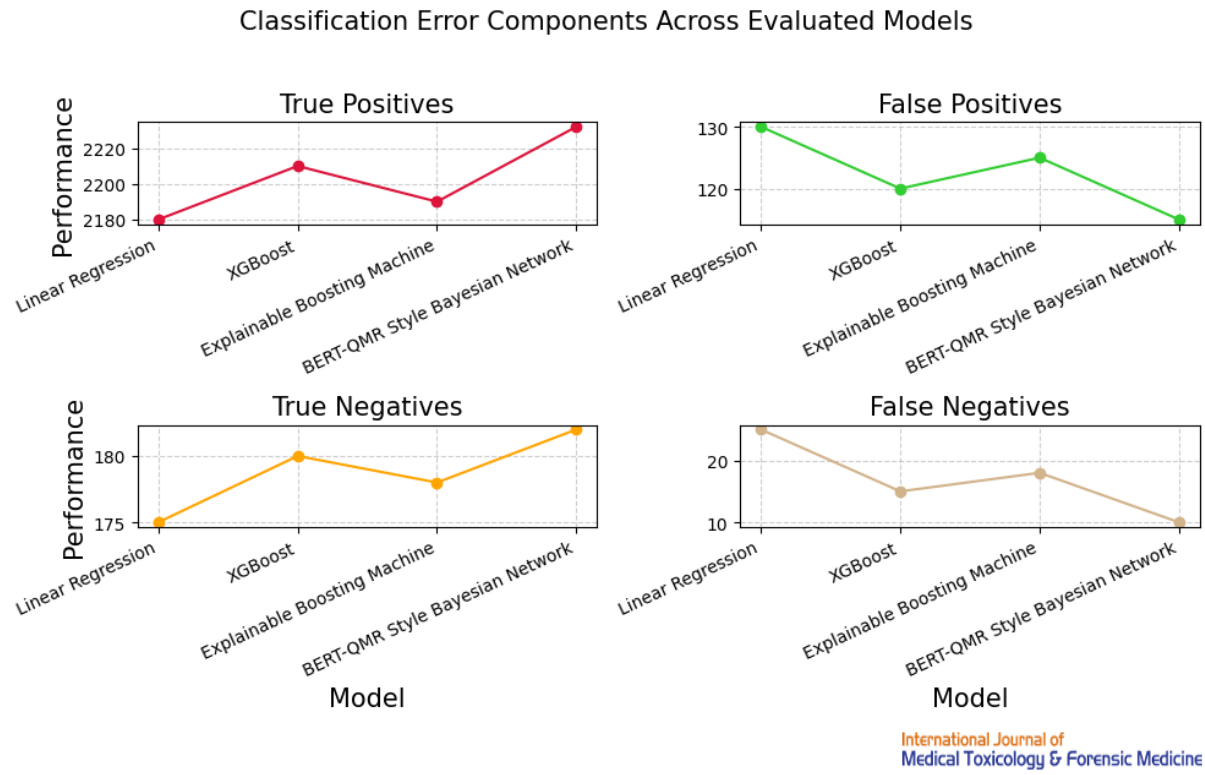


Figure 6. Comparison of classification error components across the evaluated models.

Conclusion

This study presented an AI-driven framework for mental health disorder prediction in criminal justice settings by integrating contextual language modeling with probabilistic Bayesian reasoning. The proposed BERT-QMR-style Bayesian framework combines BERT-based symptom representation with a QMR-style Bayesian network to refine initial predictions using symptom-level evidence. In addition, SMOTE, PCA, and ICA were incorporated to address class imbalance, dimensionality reduction, and behavioral and temporal feature representation. The proposed framework achieved a reported accuracy of 95.08% on the evaluated dataset and demonstrated higher reported performance than Logistic Regression, XGBoost, and Explainable Boosting Machine. The Bayesian component also provides an interpretable mechanism for updating predictions when additional symptom evidence becomes available. However, the current evaluation has important limitations. The dataset and target-label construction require clearer documentation, and the reported performance metrics require verification from the underlying predictions. Further evaluation using independent datasets, clinically validated labels, external validation, calibration analysis, and fairness assessment across relevant demographic groups is required before the framework can be considered for practical clinical or criminal-justice decision support.

Ethical Approval

Not applicable.

Acknowledgment

The authors thank all individuals who assisted with the conduct of the study.

Funding

No external funding was received for this research.

Conflicts of Interest

The authors report there are no competing interests to declare.

References

[1] Thomaidou MA, Berryessa CM. Mental illness as a sentencing determinant: A comparative case law analysis based on a machine learning approach. Crim Justice Behav. 2023;50(11):1-44. [DOI: 10.1177/00938548231170801]

[2] Prince JD, Wald C. Risk of criminal justice system involvement among people with co-

occurring severe mental illness and substance use disorder. Int J Law Psychiatry. 2018;58:1-8. [DOI: 10.1016/j.ijlp.2018.02.002]

[3] Olawade DB, Ayoola FI, Ebo TO, Asaolu AJ, Egbon E, David-Olawade AC. Artificial intelligence in forensic mental health: A review of applications and implications. J Forensic Leg Med. 2025;113:102895. [DOI: 10.1016/j.jflm.2025.102895]

[4] Altaf A, El Amraoui A, Delmotte F, Lecoutre C. Applications of artificial intelligence in mental health: A systematic literature review. J Comput Inf Syst. 2023;63(5):1280-300. [DOI: 10.1080/08874417.2022.2143455]

[5] Razavi M, et al. Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: Scoping review. JMIR Ment Health. 2024;11:e53714. [DOI: 10.2196/53714]

[6] Wagay FA, Jahiruddin. Classification of mental illnesses from Reddit posts using Sentence-BERT embeddings and neural networks. Procedia Comput Sci. 2025;258:1669-1676. [DOI: 10.1016/j.procs.2025.04.398]

[7] Nweke F, Azmee AA, Khan MAAH, Pei Y, Thomas D, Nandan M. A transformer-driven framework for multi-label behavioral health classification in police narratives. Appl Comput Intell. 2024;4(2):234-52. [DOI: 10.3934/aci.2024014]

[8] Lamba K, Rani S, Shabaz M. Explainable machine learning for mental health prediction from social media behavior: A nested cross-validation study with SHAP and LIME interpretability. Discov Ment Health. 2026;6(1):1-15. [DOI: 10.1007/s44192-026-00373-z]

[9] Dehbozorgi R, et al. The application of artificial intelligence in the field of mental health: A systematic review. BMC Psychiatry. 2025;25(1):1-18. [DOI: 10.1186/s12888-025-06483-2]

[10] Pushpa G, et al. An advanced AI framework for mental health diagnostics using Bidirectional Encoder Representations from Transformers with Gated Recurrent Units and Convolutional Neural Networks. Ing Syst Inf. 2025;30(1):213-20. [DOI: 10.18280/isi.300118]

[11] Pavez J, Allende H. A hybrid system based on Bayesian networks and deep learning for explainable mental health diagnosis. Appl Sci. 2024;14(18):8283. [DOI: 10.3390/app14188283]

[12] Kyrimi E, McLachlan S, Dube K, Neves MR, Fahmi A, Fenton N. A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. Artif Intell Med. 2021;117:102108. [DOI: 10.1016/j.artmed.2021.102108]

[13] Chatzistamatis S, et al. A fairness-aware and interpretable model for recidivism prediction. Algorithms. 2026;19(7):509. [DOI: 10.3390/a19070509]

[14] Goudarzi M, Mohammadi A, Amirpour A, Fazeli M, Nasiri MJ, Hashemi A, et al. Genetic diversity and biofilm formation analysis of Staphylococcus aureus causing urinary tract infections in Tehran, Iran. The Journal of Infection in Developing Countries. 2019;13(09):777-85. [DOI: 10.3855/jidc.11329]

[15] Farayola MM, Tal I, Saber T, Connolly R, Bendechache M. A fairness-focused approach to recidivism prediction: Implications for accuracy, trust, and equity. AI Soc. 2026;41(4):2783-801. [DOI: 10.1007/s00146-025-02452-1]

[16] Cavus M, Benli MN, Altuntas U, Sari M, Ayan H, Ugurluoglu YF. Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice. Appl Soft Comput. 2025;176:113160. [DOI: 10.1016/j.asoc.2025.113160]

[17] Bureau of Justice Statistics. Survey of Inmates in Local Jails (SILJ), 2024–2025. Washington, DC: U.S. Department of Justice; 2025. [Link]