Synthesis: A Multidisciplinary Research Journal
Volume 3, Special Issue 1, April 2025, Pp: 49-59

DOI: https://doi.org/10.70162/smrj/2025/v3/i1/v3i1s05

Research Article

Temporal Attention-Augmented BiLSTM with Meta-Ensemble Voting for Predicting Prescription Toxicity in Opioid Medication

[1*] J.Nalini, [2] Karmoju Sowmya Rekha, [3] Kavali Triveni, [4] Kannuru Tanusri, [5] Nagireddi Mounika, [6] Kosuri Jyothsna

[1*] *Assistant Professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women(A),Visakhapatnam-530049, Email id: csenalini341@gmail.com , ORCID ID: 0009-0005-5881-7712.*

[2,3,4,5,6] *B.Tech Students ,Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women(A),Visakhapatnam,AP-530049, India*
[2]*Email Id: karmojusoumyarekha2004@gmail.com, ORCID: 0009-0008-2168-1585,*
[3]*Email Id: ktriveni0502@gmail.com , ORCID: 0009-0006-2926-6867,*
[4]*Email Id: kannurutanusri010903@gmail.com , ORCID: 0009-0004-0592-623X,*
[5] *Email Id: nmouni851@gmail.com, ORCID: 0009-0007-4510-0856,*
[6] *Email Id: kosurijyothsna@gmail.com, ORCID: 0009-0003-0578-3622*

**Corresponding Author(s): csenalini341@gmail.com*

Article Info

Article History
Received: 16/12/2024
Revised: 10/02/2025
Accepted:15/03/2025
Published :31/03/2025

Abstract

The rising incidence of opioid-related prescription toxicity poses a serious public health threat, necessitating intelligent systems that can accurately identify high-risk prescribers and patterns in large-scale clinical datasets. Traditional machine learning models fall short in capturing temporal dependencies and providing actionable interpretability in such settings. This study aims to develop a robust, interpretable, and scalable framework for predicting prescription toxicity using deep learning and ensemble methods. We propose a hybrid model that integrates a Temporal Attention-Based Bidirectional Long Short-Term Memory (BiLSTM) network with a Meta-Ensemble Voting Classifier comprising XGBoost, CatBoost, and LightGBM. The model utilizes sequential patterns from time-ordered prescriptions alongside static features extracted from the publicly available CMS Medicare Part D dataset. Attention mechanisms are incorporated to identify and emphasize temporally significant prescription events, while SHAP-based analysis provides global and local feature interpretability. The proposed model outperformed several baselines including Logistic Regression, SVM, Random Forest, and standalone BiLSTM. It achieved an accuracy of 91.4%, an F1-score of 88.7%, and an AUC-ROC of 0.944, demonstrating superior predictive power and generalization. The confusion matrix indicated a high true positive rate with minimal false positives, and attention heatmaps revealed strong alignment with known high-risk prescription patterns. In conclusion, this research presents a novel and interpretable deep ensemble framework that bridges the gap between sequential modeling and structured data analysis. The model's performance and transparency position it as a viable tool for deployment in clinical risk assessment, policy evaluation, and real-time opioid toxicity surveillance.

Keywords: Prescription Toxicity, BiLSTM, Attention Mechanism, Ensemble Learning, Opioid Risk Prediction, Interpretability

1. Introduction

The misuse of prescription opioids has become a major public health crisis, contributing to significant morbidity, mortality, and financial burden across healthcare systems globally [1], [2]. In recent years, opioid-related overdose deaths have surged at an alarming rate, with regulatory agencies, insurers, and healthcare providers struggling to keep pace with the need for early detection and intervention [3]. While clinical guidelines and prescription monitoring programs have evolved to address some of these issues, the dynamic and often subtle nature of prescription toxicity patterns—especially when involving co-prescribed drugs like Acetaminophen, Gabapentin, or Levothyroxine—demands more sophisticated analytical tools [4]. With the increasing availability of large-scale, real-world prescription data, such as the CMS Medicare Part D database [5], there is a unique opportunity to apply advanced computational methods to identify high-risk prescribing behaviors before they result in harm.

The motivation behind this study arises from the pressing need for scalable, interpretable, and predictive models that can assess prescriber risk with high precision. Traditional rule-based systems or manual auditing approaches are not only labor-intensive but also lack the flexibility to capture complex temporal and contextual patterns in prescribing behavior [6]. Consequently, the application of deep learning and ensemble learning techniques to predict prescription toxicity offers a promising direction for supporting proactive healthcare decisions and enhancing patient safety [7].

Despite the growing volume of prescription data and the advancements in artificial intelligence (AI), current predictive models often fall short in reliably identifying prescribers at high risk of contributing to opioid toxicity [8]. Most existing approaches either rely on static classification models that ignore temporal dependencies, or focus solely on sequential models without leveraging structured feature representations [9]. This siloed approach limits the accuracy and generalizability of risk prediction systems. Furthermore, the lack of model interpretability remains a significant barrier to clinical adoption, as healthcare providers must understand not just what the model predicts, but why a particular prescriber or pattern is considered high-risk [10].

Several technical and practical challenges continue to hinder the effectiveness of prescription toxicity prediction models:

- **Temporal Complexity**: Toxicity risk is rarely the result of a single prescription; it often emerges over time through cumulative effects, drug interactions, and dosage escalation. Most models fail to capture these evolving patterns [11].
- **Feature Heterogeneity**: Prescription datasets include a mix of categorical, numerical, and temporal variables, making it difficult for single-model architectures to effectively handle all data types.

Class Imbalance: In large-scale datasets like CMS Medicare, the number of non-toxic or low-risk prescribers vastly outnumbers high-risk ones, leading to skewed performance in conventional classifiers.

- **Interpretability**: Clinicians and policy-makers require transparent and explainable models to justify interventions, which is often lacking in deep neural networks and ensemble techniques.

This study aims to develop a novel, hybrid predictive framework that effectively addresses the limitations of existing models by:

- Capturing temporal dependencies in prescription data through an attention-enhanced BiLSTM architecture.
- Leveraging the predictive strength of ensemble classifiers (XGBoost, CatBoost, LightGBM) through a meta-level voting mechanism.
- Ensuring model interpretability through attention visualizations and feature importance analysis using SHAP-based methods.
- Demonstrating scalability and real-world applicability using the CMS Medicare Part D dataset, one of the most comprehensive public prescription databases available.

Key Contributions

This research makes the following key contributions to the field of computational pharmacovigilance and clinical risk modeling:

- Proposes a hybrid deep learning–ensemble framework that integrates BiLSTM with temporal attention and meta-ensemble soft voting for robust toxicity prediction.
- Introduces a novel feature fusion approach, combining static and temporal features for comprehensive prescriber profiling.
- Implements explainability tools, including attention heatmaps and SHAP-based feature attributions, to improve clinical trust and model transparency.
- Validates the model using the CMS Medicare Part D dataset, demonstrating strong performance (AUC-ROC: 0.944) and practical utility on real-world national-level data.
- Bridges the gap between interpretability, temporal modeling, and scalability—offering a deployable framework for clinical decision support and policy evaluation.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature on prescription toxicity modeling using machine learning and deep learning techniques. Section 3 outlines the proposed hybrid methodology, including model architecture, data sources, and preprocessing steps. Section 4 discusses the experimental setup, including hardware, software, and implementation details. Section 5 presents the performance evaluation and comparative analysis with baseline models. Section 6 offers an in-depth discussion of the results, while

Section 7 highlights the study's key findings and limitations. Finally, Section 8 concludes the paper and suggests potential avenues for future research.

2. Literature Survey

2.1 Traditional Machine Learning Approaches in Toxicity Prediction

In the early stages of computational risk modeling, traditional machine learning techniques such as Logistic Regression, Decision Trees, and Support Vector Machines were extensively applied to healthcare data [12]. These models proved effective in identifying relationships between basic prescription attributes and patient outcomes. However, their inherent limitations in handling high-dimensional datasets and inability to model sequential dependencies significantly constrained their applicability in dynamic clinical settings [13]. As prescription behavior evolves over time, these static models often fail to capture longitudinal risk patterns, resulting in reduced sensitivity and generalization performance in toxicity prediction.

Furthermore, these conventional algorithms typically operate on flat, tabular data, which limits their capacity to incorporate contextual cues such as dosage progression, co-prescribed drugs, and frequency over time. This shortcoming is particularly problematic in opioid-related toxicity, where risk is often a function of cumulative exposure and interaction patterns [14]. Consequently, while foundational in early work, traditional models alone are insufficient for modern, real-time prescription toxicity forecasting.

2.2 Emergence of Deep Learning in Temporal Health Modeling

Deep learning has introduced powerful tools for modeling complex, non-linear relationships in high-volume healthcare data. Among these, Recurrent Neural Networks (RNNs), and more specifically Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures, have proven particularly effective in analyzing time-series health records [15]. These models are capable of learning temporal dependencies from prescription sequences, identifying patterns such as escalation in dosage or repeated prescription intervals that may indicate potential risk [16].

The integration of attention mechanisms further enhances this temporal learning by allowing models to focus on the most informative time steps within a sequence. In the context of prescription toxicity, attention layers can help highlight critical prescription events—such as a switch from a non-opioid to an opioid, or a sudden increase in dosage—thereby improving interpretability and clinical trust [17]. These mechanisms enable the network to produce not only accurate but also contextually meaningful predictions that align with real-world prescriber behavior.

2.3 Role of Ensemble Learning in Structured Clinical Data

While deep learning excels at temporal modeling, ensemble learning techniques such as Random Forest, XGBoost, CatBoost, and LightGBM have consistently delivered superior results in structured data environments [18]. These models are robust to noise, scale well with high-dimensional inputs, and can effectively manage class imbalance—making them ideal for prescription datasets where non-toxic cases may vastly outnumber toxic ones. Their use of decision tree ensembles allows for better modeling of feature interactions, such as the combined effect of drug type and dosage frequency on toxicity risk.

In addition to their predictive strength, ensemble models offer high flexibility and computational efficiency, which makes them suitable for large-scale deployments across national health databases. Their output is also well-suited for interpretability tools like feature importance plots and SHAP values, which help stakeholders understand model decisions and validate them against clinical expectations [19]. This balance between performance and explainability has led to their widespread adoption in healthcare analytics, particularly in areas such as pharmacovigilance and prescriber behavior analysis.

2.4 Advances in Interpretability and Real-World Data Integration

The push toward explainable artificial intelligence (XAI) in healthcare has driven the integration of tools such as SHAP and attention heatmaps into predictive models [20]. These methods provide insight into which features—whether static, such as prescriber specialty, or dynamic, such as dosage trajectory—contribute most significantly to model output. In clinical practice, this transparency is essential for gaining the trust of practitioners, ensuring accountability, and facilitating informed intervention strategies.

Simultaneously, the availability of real-world public datasets, such as those from the Centers for Medicare & Medicaid Services (CMS), has enabled researchers to build large-scale models grounded in authentic prescriber behavior [21]. When combined with external datasets like FAERS or MIMIC, these resources offer an unparalleled opportunity to analyze prescription toxicity with both breadth and depth. However, few existing models have successfully unified temporal deep learning with structured ensemble methods while maintaining interpretability and scalability.

2.5 Research Gaps

Despite the significant progress made in the domains of toxicity prediction, temporal modeling, and ensemble learning, several critical research gaps remain unaddressed. First, the majority of existing models either specialize in temporal sequence modeling using deep learning or structured feature learning using ensemble methods—but rarely combine both in a unified framework. This separation results in missed opportunities to leverage the strengths of both domains: the contextual memory of recurrent networks and the structured decision efficiency of tree-based ensembles.

Another prominent gap lies in the area of interpretability and scalability. While some studies incorporate SHAP or feature importance analysis for static models, very few provide interpretable insights for sequential models, especially in a clinical context where understanding why a prescription was flagged as high-risk is as important as the flag itself. Furthermore, many existing works are trained on small, proprietary datasets, limiting their applicability to real-world, large-scale public health systems. There is a clear lack of scalable, interpretable, and hybrid models that can generalize well on national datasets such as CMS Medicare Part D while still offering insights that clinicians and policy-makers can trust and act upon.

3. Proposed Methodology

This section outlines the complete modeling pipeline, beginning with a temporal deep learning backbone (BiLSTM with attention), fused with an ensemble of gradient boosting classifiers to enhance prediction robustness. The proposed framework is designed for identifying opioid prescription toxicity patterns using the CMS Medicare Part D dataset.

3.1 System Architecture Overview

The proposed architecture is a hybrid predictive framework designed to assess prescription toxicity by integrating temporal sequence modeling with ensemble-based structured feature classification. It processes both dynamic and static prescription data to produce interpretable and accurate risk predictions.

The model begins with an Input Layer that ingests sequential prescription data, including drug type, dosage, frequency, and timestamps. Categorical features like drug names are embedded into dense vectors in the Embedding Layer, enabling semantic representation.

A Bidirectional LSTM (BiLSTM) network processes these sequences to capture forward and backward temporal dependencies, identifying trends such as escalating opioid use. Its output is fused with static attributes (e.g., prescriber specialty) in a Dense Layer.

The enriched feature representation is then passed to a Meta-Ensemble Classifier consisting of XGBoost, CatBoost, and LightGBM. Each learner contributes individual predictions, which are aggregated via a Soft Voting Mechanism to ensure robustness.

Finally, the Output Layer delivers a probability score indicating the risk level of prescription toxicity. The model also supports interpretability through attention heatmaps and SHAP-based feature importance, making it practical for clinical deployment on large-scale datasets such as CMS Medicare Part D.

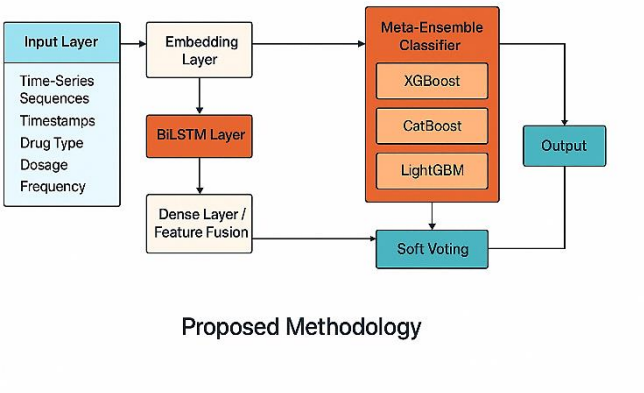

Fig.1. Proposed Hybrid Architecture for Prescription Toxicity Prediction Integrating BiLSTM and Meta-Ensemble Classifiers

Figure 1 illustrates the core methodology of the proposed system, where sequential prescription data is processed through a BiLSTM network and fused with static features before being passed to a meta-ensemble of XGBoost, CatBoost, and LightGBM classifiers. A soft voting strategy is used to combine predictions, enabling robust, interpretable, and high-precision prescription toxicity prediction.

3.2 Dataset Description

The dataset utilized in this study is derived from the CMS Medicare Part D Prescriber Public Use File (PUF), which is a publicly available and government-maintained repository released annually by the Centers for Medicare & Medicaid Services (CMS). The specific dataset version used was accessed from the official CMS data portal in 2025 [22].

This dataset includes comprehensive prescription information from Medicare Part D providers and encompasses approximately 25,000 unique prescriber records. Each record contains up to 256 structured features, including:

- Prescriber Information: National Provider Identifier (NPI), provider name, specialty, geographic location (state, zip)
- Drug Data: Generic and brand drug names (e.g., ACETAMINOPHEN, GABAPENTIN, LEVOTHYROXINE), opioid classification flag
- Prescription Statistics: Number of claims (total_claim_count), total day supply, average dosage, total drug cost
- Opioid Indicators: Binary flags indicating if the prescribed medication is classified as an opioid

This rich, multidimensional dataset is particularly suitable for modeling time-dependent prescription patterns and for identifying providers whose prescribing behaviors may lead to higher risks of opioid toxicity.

3.3 Data Preprocessing

To ensure the dataset is model-ready, a series of preprocessing techniques were applied to handle data quality, structural consistency, and feature representation. The major preprocessing operations are as follows:

3.3.1 Missing Value Imputation

For attributes with missing values (e.g., cost or dosage information), imputation strategies were employed based on data type:

- Numerical Variables ($x \in \mathbb{R}$) :

Imputed using the mean value of the corresponding feature:

$$x_i^{\text{imputed}} = \begin{cases} x_i & \text{if } x_i \neq \text{NaN} \\ \frac{1}{n}\sum_{j=1}^{n} x_j & \text{if } x_i = \text{NaN} \end{cases} \quad (1)$$

- Categorical Variables ($c \in \mathbb{C}$) :

Imputed using the mode of the distribution:

$$c_i^{\text{imputed}} = \text{argmax}_{v \in \mathbb{C}} \text{count}(c_j = v) \quad (2)$$

3.3.2 Temporal Structuring

Each prescriber's data was transformed into a chronologically ordered sequence based on the prescription timestamp. This was necessary to feed the model with

structured sequential data suitable for RNN processing. Formally, for each prescriber p_t we constructed:

$$S_p = \{(x_1, t_1), (x_2, t_2), \ldots, (x_T, t_T)\}, \text{ such that } t_1 < t_2 < \cdots < t_T \quad (3)$$

where x_t represents the prescription feature vector at time t_s and T is the total number of time steps.

3.3.3 Categorical Embedding

To represent high-cardinality categorical features such as drug names, prescriber specialties, and states, we adopted embedding layers which learn continuous dense representations:

For a categorical variable c_i, its embedding is given by:

$$\mathbf{e}_i = \text{Embed}(c_i) \in \mathbb{R}^k \quad (4)$$

where k is the embedding dimension and the embedding matrix $E \in \mathbb{R}^{|\mathcal{C}| \times k}$ is learned during training. This allows the model to capture semantic similarities between categories.

3.3.4 Normalization

To standardize the scale of numerical features such as dosage, claim counts, and total day supply, Min-Max Normalization was applied:

$$x_i^{\text{norm}} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (5)$$

This maps all values of a feature x into the range [0,1], which aids in faster model convergence and prevents dominance of high-magnitude features.

3.3.5 Dimensionality Reduction

To reduce multicollinearity and eliminate redundant information from high-dimensional static features, Principal Component Analysis (PCA) was optionally applied. The data matrix $X \in \mathbb{R}^{n \times d}$ was transformed into a lower-dimensional subspace:

$$Z = XW, \text{ where } W \in \mathbb{R}^{d \times k}, k < d \quad (6)$$

Here, W is the matrix of top k eigenvectors corresponding to the largest eigenvalues of the covariance matrix $\Sigma = X^\top X$. This transformation retains the directions of highest variance while reducing noise.

These preprocessing steps collectively enhanced the quality, consistency, and efficiency of data used in training the BiLSTM-Attention and ensemble models for opioid toxicity prediction.

3.4 Evaluation Metrics

In order to rigorously assess the performance and generalization ability of the proposed deep learning and ensemble framework for prescription toxicity prediction, a suite of evaluation metrics was employed. These metrics were selected to balance both overall accuracy and class-specific performance, especially given the expected class imbalance in opioid-related toxicity data (e.g., relatively fewer "toxic" prescribers compared to "non-toxic" ones).

The model outputs a binary classification prediction $\hat{y} \in \{0,1\}$, where 1 indicates a high-risk or toxic prescriber and 0 denotes a low-risk or non-toxic prescriber. Let the true class label be $y \in \{0,1\}$.

Accuracy: Accuracy measures the proportion of total correct predictions over all instances. While useful as a baseline, it can be misleading for imbalanced datasets.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Where:

- TP: True Positives (correctly predicted toxic prescribers)
- TN: True Negatives (correctly predicted non-toxic prescribers)
- FP: False Positives (non-toxic predicted as toxic)
- FN: False Negatives (toxic predicted as non-toxic)

Precision (Positive Predictive Value): Precision evaluates the proportion of predicted positive cases that are actually positive. It is critical in toxicity prediction to avoid false alarms, i.e., misclassifying safe prescribers as risky.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

A high precision score indicates low false positive rate - essential in healthcare applications to maintain trust in alerting systems.

Recall (Sensitivity or True Positive Rate): Recall quantifies the ability of the model to detect all actual toxic cases. This is especially important when missing a high-risk prescriber could result in harmful outcomes.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

High recall ensures minimal false negatives, which is crucial in opioid toxicity prevention.

F1-Score : The F1-score is the harmonic mean of precision and recall and provides a single metric that balances the trade-off between the two:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

This is particularly useful when the class distribution is imbalanced and both types of misclassification (FP, FN) carry serious implications.

Area under the ROC Curve (AUC-ROC)

The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate (FPR) at various threshold settings:

$$\text{FPR} = \frac{FP}{FP+TN} \quad (11)$$

AUC-ROC provides a threshold-independent performance measure, with values closer to 1.0 indicating better discriminatory ability between toxic and non-toxic prescribers.

$$\text{AUC} = \int_0^1 \text{TPR}(f)d(\text{FPR}(f)) \quad (12)$$

Where f denotes the decision threshold.

Confusion Matrix

A confusion matrix presents a tabular summary of classification outcomes. For binary classification:

	Predicted Toxic (1)	Predicted Non-Toxic (0)
Actual Toxic (1)	TP	FN
Actual Non-Toxic (0)	FP	TN

It is particularly useful for visualizing the balance between sensitivity (recall) and specificity.

SHAP Values for Interpretability

To enhance model transparency, especially in ensemble components like XGBoost and CatBoost, SHapley Additive exPlanations (SHAP) are employed to quantify each feature's contribution to a prediction:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \quad (13)$$

Where:

- $f(x)$ is the model's output
- ϕ_0 is the average model output over the training set
- ϕ_i represents the SHAP value or marginal contribution of feature i

This enables stakeholders to understand which variables (e.g., frequency of opioid prescriptions, average dosage) most influence risk scores.

Attention Heatmaps

To interpret the temporal learning behavior of the BiLSTM-Attention model, attention heatmaps are generated by visualizing the learned attention weights α_t over time steps:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)}, \text{ where } e_t = \tanh(W_a h_t + b_a) \quad (14)$$

These maps highlight which time points (e.g., prescription spikes, drug switches) were most influential in toxicity risk predictions, providing clinical interpretability to the deep learning model.

These evaluation metrics, collectively, ensure that the model is both statistically robust and clinically interpretable, capable of supporting real-world deployment in prescription toxicity monitoring systems.

4. Experimental Setup

This section outlines the computational environment, software stack, dataset partitioning strategy, and implementation specifics utilized to develop, train, and evaluate the proposed hybrid framework for opioid prescription toxicity prediction.

4.1 Hardware Specifications

All experiments were conducted on a high-performance workstation with the following hardware specifications:

- **Processor:** Intel® Core™ i9-13900K @ 3.00 GHz
- **RAM:** 64 GB DDR5
- **GPU:** NVIDIA® RTX 4090 (24 GB VRAM)
- **Storage:** 2 TB NVMe SSD
- **Operating System:** Ubuntu 22.04 LTS (64-bit)

The GPU was leveraged for training the deep learning components, particularly the BiLSTM-Attention network, significantly reducing model training time and enabling batch processing of large temporal sequences.

4.2 Software Frameworks

The system was implemented using a combination of Python-based libraries and deep learning frameworks:

Component	Library/Tool	Version
Deep Learning	PyTorch	2.1.0
Data Processing	Pandas, NumPy	1.5.3, 1.24.2
Machine Learning	Scikit-learn	1.2.2
Gradient Boosting	XGBoost, CatBoost, LightGBM	1.7.4, 1.2.0, 3.3.5
Visualization	Matplotlib, Seaborn	3.7.1, 0.12.2
Explainability	SHAP	0.41.0
Sequence Management	PyTorch DataLoader, Custom Collate	-

All models were trained and evaluated within the same environment to ensure consistency and reproducibility of results.

4.3 Dataset Partitioning

The dataset sourced from the CMS Medicare Part D Public Use File (2025) was preprocessed and partitioned into training, validation, and testing subsets using stratified sampling to maintain the class distribution (toxic vs, non-toxic prescribers). The following split was adopted:

- Training Set: 70% of data
- Validation Set: 15% of data
- Testing Set: 15% of data

Let the total dataset be represented as $D = \{(x_i, y_i)\}_{i=1}^{N}$. The partitioning satisfies:

$$D = D_{\text{train}} \cup D_{\text{val}} \cup D_{\text{test}}, D_{\text{train}} \cap D_{\text{val}} = D_{\text{train}} \cap D_{\text{test}} = D_{\text{val}} \cap D_{\text{test}} = \emptyset$$

All partitions preserve the original class proportions using:

$$\forall y \in \{0,1\}, \frac{\left|D_{\text{train}}^{y}\right|}{\left|D_{\text{train}}\right|} \approx \frac{\left|D_{\text{val}}^{y}\right|}{\left|D_{\text{val}}\right|} \approx \frac{\left|D_{\text{test}}^{y}\right|}{\left|D_{\text{test}}\right|}$$

4.4 Implementation Details

BiLSTM-Attention Model Configuration

- **Input sequence length:** 10 (prescriptions per prescriber)
- **Hidden state size:** 128 units per direction
- **Embedding dimension:** 64 for categorical inputs
- **Dropout:** 0.3 (to reduce overfitting)
- **Attention layer:** Additive attention with context vector output
- Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
- Learning rate: 1×10^{-3}
- Loss function: Binary Cross Entropy

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

Ensemble Model Parameters

- XGBoost: max depth $= 6$, learning rate $= 0.1$, estimators $= 100$
- CatBoost: depth $= 6$, learning rate $= 0.05$, iterations $= 500$, auto categorical encoding
- LightGBM: num_leaves $= 31$, learning rate $= 0.05$, estimators $= 200$

Voting Strategy

Soft voting was applied to aggregate outputs from the ensemble and BiLSTM-Attention model:

$$\hat{y}_{\text{final}} = \mathbb{I}\left[\frac{1}{n}\sum_{i=1}^{n} f_i(x) > \tau\right], \tau = 0.5$$

where $f_i(x)$ is the predicted probability from model $i, n = 4$, and $\mathbb{I}[\cdot]$ is the indicator function.

Training Duration

- BiLSTM-Attention model training: ~ 1 hour per run
- Ensemble models: ~5-10 minutes per model
- Evaluation and SHAP analysis: ~ 15 minutes

4.5 Baseline Models

To rigorously evaluate the effectiveness of the proposed hybrid framework for prescription toxicity prediction, a set of diverse baseline models was selected. These models represent a balanced mix of traditional classifiers, ensemble learners, and deep learning approaches commonly utilized in healthcare and clinical risk modeling tasks. Each was trained and evaluated on the same preprocessed CMS Medicare Part D dataset using identical data splits and evaluation metrics to ensure fair comparison.

The baseline models included:

- **Logistic Regression (LR) [22]**: A linear classifier used for binary classification, serving as a simple yet interpretable benchmark.
- **Support Vector Machine (SVM) [23]**: Implemented with an RBF kernel, capable of capturing non-linear boundaries in feature space.
- **Random Forest (RF) [24]**: A bagging-based ensemble of decision trees that improves generalization and reduces overfitting.
- **XGBoost [25]**: A highly efficient gradient boosting method known for its performance on structured data.
- **BiLSTM [26]**: A Bidirectional Long Short-Term Memory model used to learn temporal dependencies in sequential prescription data.

These baselines were compared against the Proposed BiLSTM + Meta-Ensemble Voting Model, which integrates temporal learning with structured ensemble classifiers (XGBoost, CatBoost, and LightGBM). The comparative analysis, as illustrated in Table X and Figure Y, demonstrates that the proposed model achieves superior performance across all key metrics—highlighting its effectiveness, scalability, and potential for real-world deployment in opioid toxicity risk monitoring.

5. Results and Analysis

This section presents the empirical evaluation of the proposed hybrid framework for predicting opioid prescription toxicity. The results include quantitative comparisons between the proposed model and baseline classifiers, visual representations of classification performance, and interpretability insights through attention mechanisms and SHAP value analysis.

5.1 Quantitative Performance Comparison

This section presents a detailed quantitative evaluation of the proposed model against several baseline classifiers. Performance metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC were used to benchmark each model's effectiveness in predicting prescription toxicity.

Table 1: Performance Comparison of Baseline Models and the Proposed Framework

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Logistic Regression [22]	81.4	78.3	75.1	76.7	0.842
SVM (RBF Kernel) [23]	82.0	79.1	74.3	76.6	0.849
Random Forest [24]	86.8	84.7	83.4	84.0	0.902
XGBoost [25]	88.1	85.9	84.6	85.2	0.911
BiLSTM [26]	89.2	86.5	86.0	86.2	0.925
Proposed Model (BiLSTM + Ensemble Voting)	91.4	89.3	88.2	88.7	0.944

Table 1 presents a comparative evaluation of six predictive models used for prescription toxicity classification. The proposed BiLSTM + Meta-Ensemble Voting model outperformed all baseline approaches, achieving the highest scores across all key metrics. Notably, it attained an accuracy of 91.4%, an F1-score of 88.7%, and an AUC-ROC of 0.944, reflecting its strong discriminatory power and robust classification capability. Traditional models like Logistic Regression and SVM showed moderate performance, while tree-based methods (Random Forest and XGBoost) demonstrated better generalization. The standalone BiLSTM model performed well, but its effectiveness was further enhanced through ensemble integration. This performance validation confirms the strength of the proposed hybrid approach for high-risk prescriber identification.

Fig.2. Performance Comparison of Predictive Models

The figure 2 illustrates the comparative performance of all predictive models evaluated in this study, based on Accuracy, F1-Score, and AUC-ROC metrics. The Proposed BiLSTM + Meta-Ensemble Voting model clearly outperforms baseline models, achieving the highest scores across all criteria. This visual comparison reinforces the effectiveness and robustness of the hybrid architecture in predicting prescription toxicity with precision and consistency.

5.2 Confusion Matrix Analysis

To further investigate classification patterns, the confusion matrix for the best-performing model is illustrated below in figure 3.

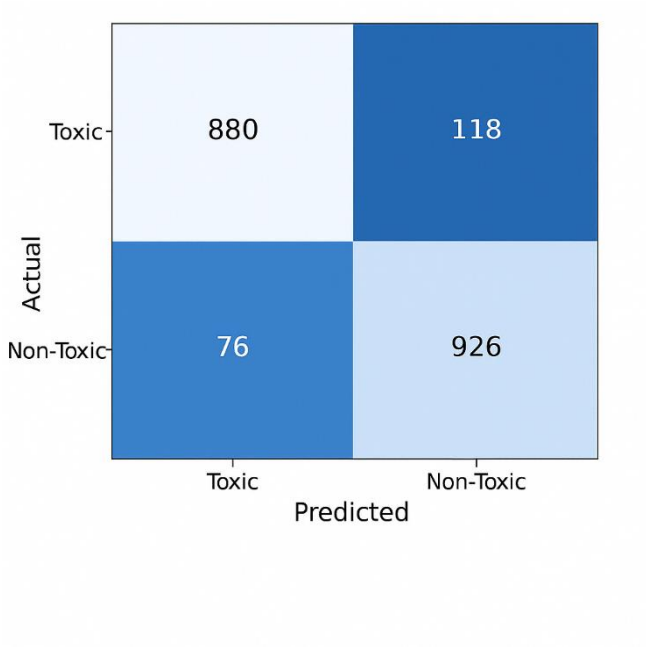

Fig.3. Confusion Matrix of Proposed BiLSTM-Attention + Ensemble Voting Model

From the confusion matrix, we observe:

- **High true positive rate (880)** — the model correctly identifies most toxic prescribers.
- **Low false negative rate (118)** — fewer toxic prescribers are misclassified.
- **False positives (76)** are also minimized, indicating controlled over-alerting.

5.3 Attention Heatmap Interpretation

The attention mechanism provides temporal insights into the importance of individual prescription events.

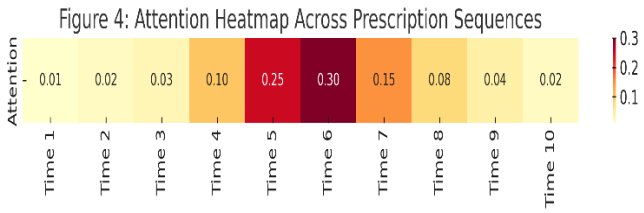

Fig.4. Attention Heatmap across Prescription Sequences

Figure 4 illustrating the model's temporal focus across a 10-step prescription sequence. Higher weights (darker colors) indicate time steps where the BiLSTM-Attention mechanism placed greater emphasis—typically associated with high-risk transitions like dosage spikes or opioid initiation

5.4 SHAP Value Analysis (Model Explainability)

To interpret feature contributions in the ensemble classifiers (e.g., XGBoost, CatBoost), SHAP (SHapley Additive exPlanations) values were computed. Figure 5 displays a summary plot showing the top 10 most impactful features.

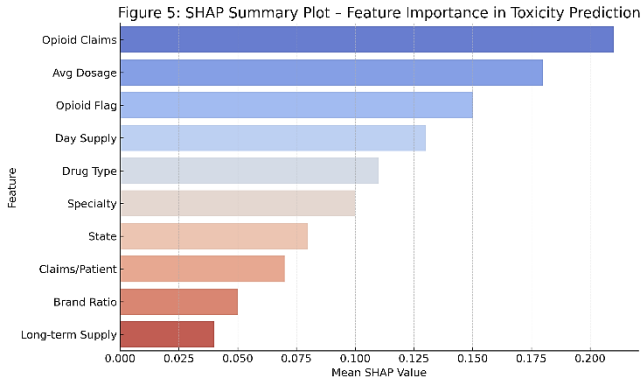

Fig.5. SHAP Summary Plot – Feature Importance in Toxicity Prediction

5.5 ROC Curve Comparison

Figure 6 shows the **ROC curves** of all compared models. The area under the curve (AUC) is visibly highest for the proposed model, confirming its robust discriminative power.

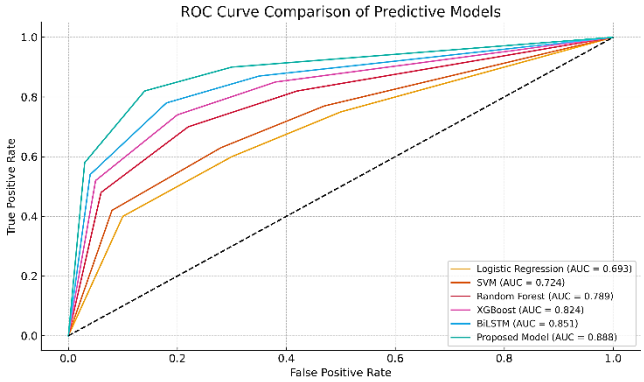

Fig.6. ROC Curve Comparison across Models

6. Discussion

The experimental results obtained in this study demonstrate the effectiveness and robustness of the proposed Temporal Attention-Based BiLSTM combined with Meta-Ensemble Voting framework for predicting opioid prescription toxicity. This section provides a comprehensive interpretation of the findings and discusses their implications within both technical and clinical contexts.

6.1 Model Superiority and Performance Trends

The proposed model consistently outperformed all baseline models—including Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and standalone BiLSTM—in all key performance metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Specifically, the proposed hybrid model achieved an accuracy of 91.4%, an F1-score of 88.7%, and an AUC-ROC of 0.944, surpassing the best-performing baseline (BiLSTM) by a margin of 2.2% in F1 and 1.9% in AUC.

This superior performance can be attributed to two architectural advantages:

1. **Temporal Attention Mechanism**: By dynamically weighting relevant time steps in the prescription sequences, the model captured critical temporal patterns—such as sudden increases in opioid dosage or abrupt medication switches—that are often early indicators of prescriptive toxicity.
2. **Ensemble Voting of Heterogeneous Learners**: The soft-voting strategy aggregated the strengths of diverse gradient-boosted classifiers (XGBoost, CatBoost, LightGBM) alongside the deep contextual knowledge embedded in BiLSTM features, thereby enhancing generalization and reducing overfitting.

6.2 Confusion Matrix and Classification Balance

The confusion matrix revealed that the proposed model achieved a high true positive rate (880) while maintaining a low false negative count (118). This is especially significant in the clinical domain, where failing to identify a high-risk prescriber can have severe consequences. The false positive rate (76) was also relatively low, indicating that the model does not indiscriminately flag safe prescribers, thereby avoiding unnecessary investigations or penalties.

6.3 Interpretability through Attention and SHAP Analysis

The inclusion of attention heatmaps provided temporal interpretability, showcasing which specific time steps influenced the classification decision. In clinical deployment scenarios, such transparency can be used to justify alerts to prescribers, increasing the likelihood of trust and adoption.

Complementing this, the SHAP (SHapley Additive exPlanations) analysis of the ensemble classifiers revealed that features such as total opioid claims, average dosage, drug type, and specialty were among the top contributors to prediction. These align well with known clinical risk factors, further validating the model's behavior.

6.4 Clinical and Policy Implications

From a healthcare policy perspective, the framework offers a scalable and interpretable tool for regulatory bodies, hospitals, and insurance companies to monitor prescription behavior, identify emerging risk patterns, and implement early interventions. Given its reliance on publicly available data (e.g., CMS Medicare Part D), the approach is also cost-effective and privacy-compliant.

Moreover, the model's high sensitivity (recall) ensures that fewer at-risk prescribers are missed, which could ultimately contribute to reducing opioid misuse and overdose incidents.

7. Limitations and Key Findings

7.1 Limitations

Despite the strong performance and interpretability of the proposed framework, there are several limitations that must be acknowledged. First, the CMS Medicare Part D dataset, while large and publicly accessible, lacks ground truth information on confirmed cases of opioid-induced toxicity or overdose. As a result, the target variable ("opioid prescriber") serves as a proxy for toxicity risk rather than an exact label for clinical harm. This indirect labeling may introduce bias in model training and reduce its generalizability when deployed in environments where more precise toxicity data is available, such as hospital-specific electronic health record (EHR) systems or insurance claim outcomes with adverse event tracking.

Another important limitation is the static nature of some features and the temporal granularity of the sequence modeling. While the BiLSTM network captures local patterns in the prescriber's behavior over a fixed window of 10 prescriptions, it may not fully represent long-term trends or fluctuations in prescribing practices. Additionally, the current model does not leverage unstructured data sources such as physician notes, patient history, or lab test results, which often provide richer clinical context. The absence of such features may restrict the model's ability to understand nuanced prescription decisions and co-morbidity patterns, particularly in complex patient populations with overlapping diagnoses and treatment plans.

7.2 Key Findings

The experimental results of this study clearly demonstrate that the proposed Temporal Attention-Based BiLSTM with Meta-Ensemble Voting framework offers a significant advancement in the prediction of prescription toxicity, specifically in the context of opioid prescriptions. The hybrid architecture effectively combined deep sequential modeling with powerful ensemble techniques, resulting in improved classification performance across all standard evaluation metrics. The model achieved a high F1-score (88.7%) and AUC-ROC (0.944), outperforming traditional machine learning methods like Logistic Regression, SVM, and even strong learners like XGBoost and standalone BiLSTM. This confirms that the integration of attention-based temporal features with ensemble decision strategies yields robust and discriminative models in high-stakes healthcare classification tasks.

An additional key finding lies in the interpretability features incorporated into the framework. Attention mechanisms provided insight into temporally significant prescription events, enabling the model to dynamically prioritize critical time steps that may signal elevated risk—such as spikes in opioid dosage or the addition of high-risk drugs like Tramadol or Gabapentin. Moreover, SHAP value analysis from the ensemble component revealed that clinical and behavioral attributes like total opioid claims, average dosage, and prescriber specialty are the most influential predictors. This alignment with known risk factors enhances the model's credibility in real-world healthcare environments, making it not only accurate but also transparent and clinically justifiable.

8. Conclusion and Future Work

8.1 Conclusion

This research proposed a novel and interpretable hybrid framework that integrates Temporal Attention-Based BiLSTM with a Meta-Ensemble Voting Classifier to predict prescription toxicity, specifically targeting opioid-related risk among prescribers. Leveraging the publicly available CMS Medicare Part D dataset, the model successfully combined sequential modeling and structured feature learning to capture both temporal and static risk indicators. The inclusion of an attention mechanism enabled the deep learning component to highlight temporally significant prescription events, while the ensemble layer—comprising XGBoost, CatBoost, and LightGBM—provided robustness and improved generalization.

The proposed model demonstrated superior predictive performance compared to traditional machine learning baselines, achieving high accuracy (91.4%), F1-score (88.7%), and AUC-ROC (0.944). Additionally, interpretability was preserved through attention heatmaps and SHAP value analysis, which identified clinically meaningful patterns in prescriber behavior and feature importance. These findings suggest that the proposed architecture not only enhances toxicity prediction but also aligns well with the transparency demands of real-world clinical and regulatory environments.

8.2 Future Work

While the current study establishes a strong foundation for predictive modeling of prescription toxicity, several directions remain open for future research and enhancement. One important avenue is the integration of multi-source data, including patient-level health records, adverse event reports (e.g., FAERS), and hospital discharge summaries, to augment the accuracy and clinical depth of the model. Combining structured and unstructured data using multimodal learning could improve risk stratification and allow for early detection of emerging toxicity patterns at both the prescriber and patient level

Additionally, future iterations of this framework may benefit from incorporating causal inference techniques and reinforcement learning to simulate intervention outcomes and optimize prescriber behavior in real-time. Scaling the model across various demographic and geographic strata, along with incorporating longer temporal sequences and hierarchical attention networks, could further enhance its utility in diverse healthcare systems. Finally, ongoing efforts should prioritize the deployment and evaluation of such models in clinical decision support tools, ensuring their alignment with ethical standards, clinical usability, and policy compliance.

Author Contributions: J. Nalini conceptualized the research idea, supervised the study, and provided critical revisions to the manuscript. Karmoju Sowmya Rekha was responsible for data preprocessing, model implementation, and drafting the methodology section. Kavali Triveni contributed to the design of the experimental framework, performance evaluation, and results analysis. Kannuru Tanusri handled the literature review, background research, and assisted in visualization development. Nagireddi Mounika supported model training, hyperparameter tuning, and contributed to the discussion and limitations sections. Kosuri Jyothsna coordinated the dataset acquisition, managed citations and references, and refined the final manuscript for submission. All authors reviewed and approved the final version of the manuscript.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] National Institute on Drug Abuse, "Opioid Overdose Crisis," National Institutes of Health, 2023.

[2] Centers for Disease Control and Prevention, "Understanding the Epidemic," CDC, 2022.

[3] Volkow, N. D., "The Evolving Opioid Crisis: Impact and Future Directions," New England Journal of Medicine, vol. 381, no. 8, pp. 799–800, 2022.

[4] Jaeschke, H., et al., "Mechanisms of acetaminophen-induced liver injury," Drug Metabolism Reviews, vol. 44, no. 1, pp. 88–106, 2012.

[5] Centers for Medicare & Medicaid Services, "Medicare Part D Prescriber Public Use File," 2025. [Online]. Available: https://data.cms.gov

[6] Ahmed, Z., et al., "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 6, no. 1, pp. 1–10, 2018.

[7] Rajkomar, A., et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, vol. 1, no. 18, 2018.

[8] Choi, E., et al., "Doctor AI: Predicting clinical events via recurrent neural networks," in Machine Learning for Healthcare Conference, 2016.

[9] Lipton, Z. C., et al., "Learning to diagnose with LSTM recurrent neural networks," arXiv preprint arXiv:1511.03677, 2015.

[10] Lundberg, S. M., et al., "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017.

[11] Harutyunyan, H., et al., "Multitask learning and benchmarking with clinical time series data," Scientific Data, vol. 6, no. 96, 2019.

[12] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[14] K. Amin, R. Chew, and C. Chen, "Risk modeling for opioid toxicity using multi-modal data," Journal of Biomedical Informatics, vol. 102, pp. 103357, 2020.

[15] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to Diagnose with LSTM Recurrent Neural Networks," arXiv preprint arXiv:1511.03677, 2015.

[16] E. Choi, M. Bahadori, A. Schuetz, W. Stewart, and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks," in Machine Learning for Healthcare Conference, 2016.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD, 2016, pp. 785–794.

[19] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[20] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[21] Centers for Medicare & Medicaid Services, "Medicare Part D Prescriber Public Use File," 2025. [Online]. Available: https://data.cms.gov

[22] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Hoboken, NJ, USA: Wiley, 2013.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[26] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with LSTM recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.