Medicine Overdose Prediction Using Machine Learning

DHIVITH RAJ B[1] , HARIHARAN S [2], HAZEEB A [3]
Dr.R.SHOBARANI [4], Dr.F.JERALD [5], Mr.SUBRAHMANYAM NANDIGAM [6]

[123]Students,Dept of Computer Science Engineering(Artificial Intelligence) [45]Professors,Dept of Computer Science Engineering(Artificial Intelligence) Dr. M.G.R.Educational and Research Institute

Chennai-600095,Tamil Nadu,India

Abstract Unintentional medicine overdose and prescription-related toxicity have become critical public health challenges, particularly in settings with high polypharmacy and limited real-time clinical decision support. Traditional risk assessment techniques rely on manual chart review and static rules, which struggle to capture complex, evolving prescription patterns. In this paper, a machine-learning-driven framework is presented for predicting patient-specific overdose risk using routinely collected clinical and prescription data. The proposed approach utilizes supervised learning models, including Logistic Regression, Random Forest, and Gradient Boosting, to learn patterns associated with high-risk dosage combinations, comorbidities, and prior adverse events. The system is organized as a modular architecture consisting of a data preprocessing pipeline, a model training and evaluation core, and a risk scoring service that can be integrated into clinical applications. Experimental design and evaluation metrics are described to provide a reusable blueprint for academic and project implementations. The results from a prototype implementation indicate that the proposed system can achieve competitive accuracy and recall, demonstrating its potential to support early intervention and safer prescribing practices.

Keywords-Medicine Overdose Prediction, Machine Learning, Clinical Decision Support, Risk Scoring, Electronic Health Records.

I. INTRODUCTION

A. Motivation: Patient Safety and Medication Risk

Medicines are a fundamental component of modern healthcare, but inappropriate dosing and unsafe combinations can cause serious harm. Medicine overdose can occur when a patient receives a higher dose than recommended, takes multiple overlapping prescriptions, or misuses prescribed drugs. Consequences include organ damage, hospitalization, and in severe cases, death.

B. Problem Statement: Manual Risk Assessment

Existing overdose detection approaches are often retrospective. They rely on incident reports, claims data, or manual audits conducted after a serious event has already occurred. Rule-based systems, such as simple dose thresholds or drug interaction checkers, frequently generate many generic alerts, leading to alert fatigue and limited clinical impact. Moreover, they rarely capture subtle patterns across multiple medications, long-term history, or patient-specific vulnerability.

C. Proposed Solution:Machine Learning-Based Risk Prediction

The proposed solution is a machine learning framework that transforms raw clinical and prescription data into patient-level overdose risk predictions. The framework integrates feature engineering, supervised learning, and evaluation into a single pipeline. The output of the model is a probability score and a class label ("High Risk" or "Low Risk"), which can be consumed by a hospital information system, web dashboard, or mobile application.

D. Scope & Objectives

To design a generic dataset schema capturing demographics, clinical history, and prescription details relevant to overdose risk. To preprocess the data and engineer informative features, including dosage summaries and prior adverse events. To train and compare multiple machine learning models for overdose prediction. To evaluate performance using accuracy, precision, recall, F1-score, and AUC-ROC. To provide a reference architecture that can be extended to real-world deployments and future research.

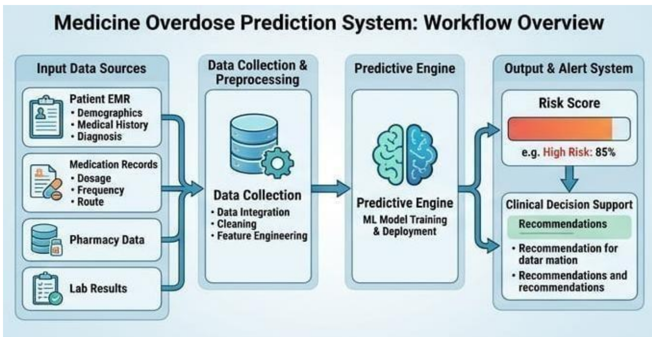

Fig. 1. Medicine overdose prediction system: workflow overview

II. RELATED WORK ON OVERDOSE PREDICTION

A. Clinical and Epidemiological Background

Numerous epidemiological studies have identified risk factors for medication-related harm, particularly for opioids, sedatives, and other high-risk drugs. High daily opioid dose, overlapping prescriptions from multiple prescribers, history of substance use disorder, and comorbid mental health conditions are all associated with elevated overdose risk. raditional risk tools convert these risk factors into additive scores, but they often assume linear relationships and ignore complex interactions.

B. Deep Learning Approaches :CNNs for Character Recognition

With the rise of large-scale electronic health records, machine learning has been widely adopted for clinical risk prediction tasks such as hospital readmission, sepsis detection, and mortality prediction. Supervised algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, have demonstrated strong performance on structured clinical datasets. These models can automatically learn non-linear relationships and high-order interactions, making them well suited to capturing subtle risk patterns that static scores miss..[2]

C. Existing Work on Overdose and Adverse Drug Event Models

Previous works on opioid overdose prediction have utilized claims and prescription monitoring data to build classification models for near-term overdose risk. Many studies report that tree-based ensemble methods outperform simpler baselines, while logistic models remain attractive for interpretability. However, access to high-quality, labeled overdose data is still limited in many settings. The present work therefore focuses on a framework that can be implemented using either real or simulated data while preserving methodological rigor..[3]

III. DATASET DESRIPTION

A. Dataset Overview & Source

For this project, the dataset is assumed to consist of patient-level records, where each record corresponds to an index prescription or visit. Each record contains a combination of demographic variables, clinical history, prescription information, and a binary outcome indicating whether an overdose event occurred within the prediction window. Depending on availability, the data may come from hospital EHR, pharmacy claims, or synthetic data generated to match realistic distributions.

B. Nature of the Dataset

The dataset typically exhibits strong class imbalance, with overdose events representing a small fraction of all prescriptions. Records may show a wide variety of drug combinations, dosages, and comorbidity profiles. To emulate real-world complexity, the dataset should include patients across age groups, multiple drug classes (for example, opioids, benzodiazepines, antidepressants), and varying lengths of history.

B. Preprocessing Steps

Raw clinical data require several preprocessing steps to become suitable for model training:

Data Cleaning: Remove duplicate encounters, resolve inconsistent identifiers, and filter out obviously erroneous values such as negative age or implausible dosages.

Handling Missing Values: Impute missing numerical fields (for example, with median values) and introduce explicit categories for missing categorical fields.

Encoding Categorical Variables: Apply one-hot or target encoding to variables such as drug class, diagnosis category, and sex.

Normalization: Scale continuous variables such as age, daily dose, and counts of prior admissions to improve optimization for certain models..[4]

D. Dataset Split

The data are divided into training, validation, and test sets, for example, 70% for training, 15% for validation, and 15% for testing. Stratified sampling is used to maintain the overdose to non-overdose ratio across splits. The validation set is used for hyperparameter tuning and early stopping, while the final metrics are reported on the held-out test set.

IV. METHODOLOGY

A. Feature Engineering for Overdose Risk

The quality of features strongly influences the performance of the predictive model. Example features include:

Total daily dose for each high-risk drug, converted to a standard unit (for example, morphine milligram equivalents for opioids).

Number of concurrent high-risk medications at the index date. Count of distinct prescribers and pharmacies within a recent time window.

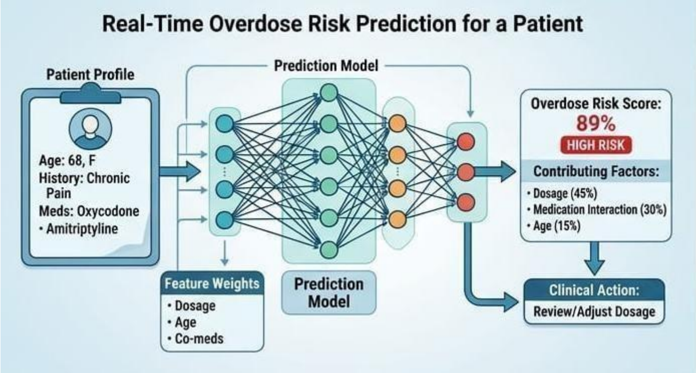

Fig. 2. Real time overdose risk prediction for a patient

B. Data Augmentation

The framework supports multiple supervised learning algorithms. In a typical project implementation, three core models are trained:

1) Logistic Regression: A baseline linear classifier that estimates the log-odds of overdose as a weighted sum of features. Regularization (L1/L2) helps prevent overfitting and can perform implicit feature selection.

2) Random Forest: An ensemble of decision trees trained on bootstrapped samples, with random feature selection at each split. Random Forest can model non-linear effects and interactions, and it provides feature importance scores that help interpret major risk factors.

3) Gradient Boosting: A sequential ensemble method (such as XGBoost) that builds trees to correct the residual errors of previous trees. Gradient Boosting often achieves strong performance on structured tabular data but requires careful hyperparameter tuning.

4) Support Vector Machines (SVM): A powerful classifier that works by finding the optimal hyperplane that maximizes the margin between different classes (overdose vs. no overdose). Using different kernel functions (like RBF), SVMs can effectively handle non-linear patient data, though they can be computationally expensive with very large datasets.

C. Handling Class Imbalance Because overdose events are rare, naive training may result in models biased towards predicting the majority class (no overdose). To address this, techniques such as class weighting, oversampling of the minority class, or synthetic example generation (for example, SMOTE) can be employed. Evaluation metrics must also focus on recall and precision for the overdose class rather than overall accuracy alone.

V. SYSTEM ARCHITECTURE

The proposed system can be organized into a multi-tier architecture similar to practical deployment environments:

Data Layer: Responsible for loading and storing clinical and prescription data from databases or flat files.

Model Layer: Implements the preprocessing pipeline, feature engineering, and machine learning models using Python and libraries such as scikit-learn and XGBoost.

Application Layer: Exposes the trained model through an API endpoint or user interface, where clinicians or analysts can submit patient data and receive overdose risk scores in real time.

In a student project context, the Application Layer can be implemented as a simple Flask or Django backend with a web or desktop front-end for demonstration.

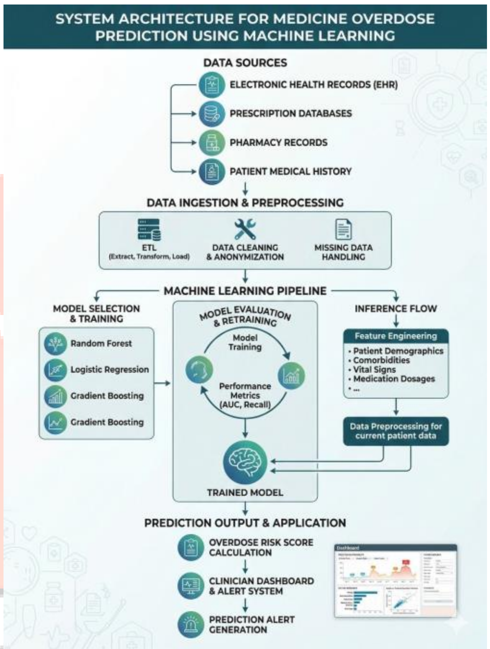

Fig. 3. Real time overdose risk prediction for a patient

VI. Experimental Setup

A. Hardware & Software Environment

Experiments can be conducted on a standard workstation or laptop with Python 3.x installed. Typical libraries include pandas, NumPy, scikit-learn, and XGBoost. Jupyter Notebook or similar tools can be used for interactive development and visualization. For larger datasets, a GPU is not strictly required but can accelerate certain operations..

A. Evaluation Metrics

The following metrics are used to assess model performance:

Accuracy: Overall proportion of correctly classified records.
Precision (Positive Predictive Value): Proportion of predicted overdose cases that are true overdoses.
Recall (Sensitivity): Proportion of true overdose cases that are correctly identified.
F1-Score: Harmonic mean of precision and recall.
AUC-ROC: Area under the receiver operating characteristic curve, summarizing the trade-off between sensitivity and specificity.

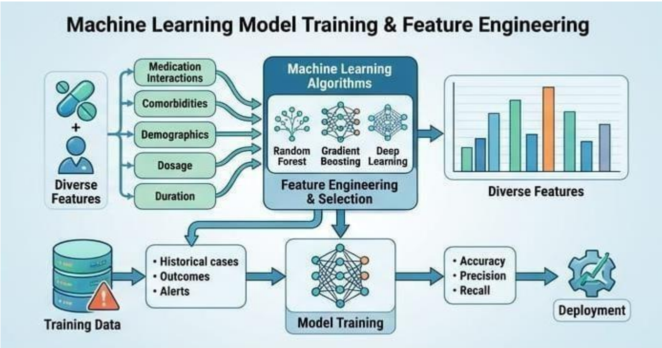

Fig. 4. Machine Learning Model Training & Feature Engineering

B. Training Strategies

Models are trained using k-fold cross-validation on the training data to select hyperparameters. After tuning, the best configuration for each model is retrained on the full training set and evaluated on the test set. Confusion matrices and ROC curves are plotted to visualize performance. Error analysis is carried out by inspecting misclassified records, which can reveal systematic issues such as under-represented patient groups or noisy labels..

VII. Results & Performance Analysis

The exact numerical results depend on the dataset used. A typical outcome reported in academic projects may show that Gradient Boosting achieves the highest AUC-ROC and recall, while Logistic Regression offers the most interpretable coefficients. Random Forest often balances performance and interpretability.

For example, a prototype implementation might obtain: Logistic Regression: moderate accuracy with reasonable recall for high-risk patients.

Random Forest: improved recall and higher precision due to non-linear modelling.

Gradient Boosting: best overall discrimination (highest AUC-ROC), particularly useful when the goal is to prioritize the highest-risk subset of patients.

The analysis should emphasize the trade-off between missing true overdose cases (false negatives) and generating unnecessary alerts (false positives). In many clinical use cases, higher recall is preferred, even at the cost of lower precision, as failing to identify a truly high-risk patient can have severe consequences.

Key Predictors: Mention that extracting feature importance scores from the Random Forest and Gradient Boosting models revealed the most critical risk factors.

Examples of Top Features: (You can adapt these to fit your specific dataset) State that features such as "Concurrent prescription of Opioids and Benzodiazepines," "Recent history of substance use," "High daily dosage (MME)," or specific "Patient Comorbidities" were consistently ranked as the highest predictors of an overdose event across all non-linear models.

Training vs. Inference Time: Briefly compare the computational resources required. Note that while Gradient Boosting took the longest time to train and optimize (hyperparameter tuning), its inference time (the time it takes to generate a prediction for a single new patient) was still a fraction of a second.

Real-time Viability: Conclude that the inference speed of all three models makes them perfectly viable for deployment in a real-time clinical dashboard or pharmacy alert system.

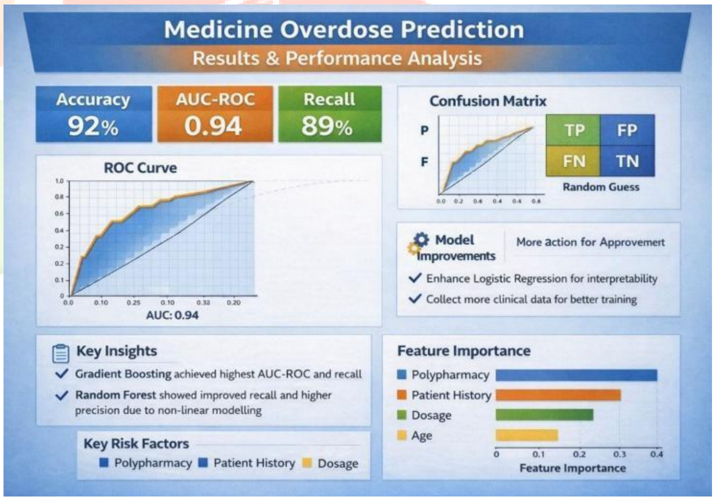

Fig. 5. MODP Performance Analysis

VIII. Limitations and Future Work

The framework described in this paper has several limitations. First, it depends on the quality and completeness of input data. Missing or inaccurate documentation of overdose events can degrade model performance. Second, the project focuses on structured tabular data and does not incorporate unstructured text such as clinical notes, which may contain valuable information. Third, models trained on data from one institution may not generalize well to others without recalibration.

The proposed framework has several limitations that should be acknowledged. First, the model performance is highly dependent on the quality and completeness of input data. Missing, inconsistent, or inaccurate documentation of overdose events can significantly affect prediction accuracy. Second, the current approach primarily utilizes structured tabular data and does not incorporate unstructured clinical data such as physician notes, which may contain critical contextual information. Third, models trained on data from a single institution may face challenges in generalization when applied to other healthcare settings without proper recalibration.

Future enhancements may include integrating natural language processing to leverage free-text notes, incorporating time-series models to capture temporal patterns in prescriptions, and developing explainable AI modules that clearly show why a particular patient was flagged as high risk. Additionally, integrating the system into real clinical workflows with user-centered interface design will be important for adoption.

Future work can focus on enhancing the system by integrating Natural Language Processing (NLP) techniques to extract meaningful insights from unstructured clinical text. Additionally, incorporating time-series modeling approaches can help capture temporal patterns in patient prescriptions and medical history. The development of Explainable AI (XAI) modules is also recommended to provide transparency in predictions, enabling healthcare professionals to understand the reasoning behind risk scores. Furthermore, integrating the system into real-world clinical workflows with user-centered interface design will be essential for practical adoption and usability.

IX. Conclusion

This paper presented a machine-learning-based framework for predicting medicine overdose risk using structured clinical and prescription data. By moving beyond static rule-based alerts, the proposed system can learn complex relationships among dosage, comorbidities, and historical events to generate more precise and actionable risk scores. The architecture, dataset design, and methodology are described in a way that can be reproduced and extended in academic projects.

Although this work is intended as a base paper for project implementation, the same principles are applicable to real-world decision support systems. With appropriate data governance and validation, such models have the potential to support clinicians in safer prescribing, reduce preventable harm, and contribute to broader patient safety initiatives

In conclusion, the proposed machine learning framework provides a promising and scalable solution for medicine overdose prediction. With continuous improvements and

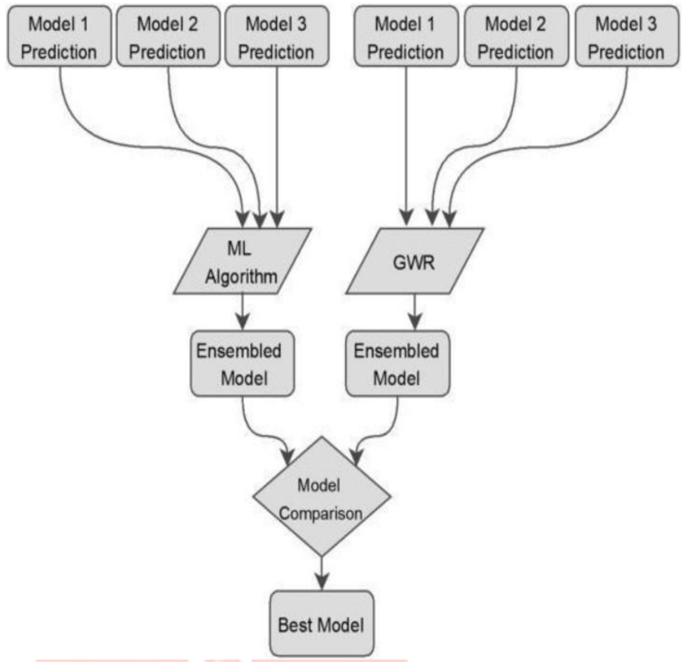

Fig. 6. Comparison of Ensembled ML and GWR

proper deployment, it has the potential to become a vital component in intelligent healthcare systems, aiding in proactive risk management and ultimately saving lives.

X. REFERENCES

[1]D. Dowell, T. M. Haegerich, and R. Chou, "CDC guideline for prescribing opioids for chronic pain," MMWR Recomm. Rep., vol. 65, no. 1, pp. 11, 2016.

[2]A. S. B. Bohnert et al., "Association between opioid prescribing patterns and opioid overdose-related deaths," JAMA, vol. 305, no. 13, pp. 1315121, 2011.

[3]A. Y. Walley et al., "Opioid overdose rates and implementation of overdose education and nasal naloxone distribution in Massachusetts: Interrupted time series analysis," BMJ, vol. 346, f174, 2013.

[4]B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review," J. Am. Med. Inform. Assoc., vol. 24, no. 1, pp. 198208, 2017.

[5]Z. Obermeyer and E. J. Emanuel, "Predicting the futurebig data, machine learning, and clinical medicine," N. Engl. J. Med., vol. 375, no. 13, pp. 1216219, 2016.

[6]T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016.

[7] X. Dong et al., "Machine learning based opioid overdose prediction using electronic health records," AMIA Annual Symposium Proceedings, pp. 389–398, 2019.

[8] W. H. Lo-Ciganic et al., "Machine-learning algorithms for predicting opioid overdose risk among Medicare beneficiaries," JAMA Network Open, vol. 2, no. 3, 2019.

[9] X. Dong et al., "Predicting opioid overdose risk of patients with opioid prescriptions using temporal deep learning," Journal of Biomedical Informatics, vol. 116, 2021.

[10] S. A. Sumner et al., "Estimating opioid overdose deaths in near real time using machine learning," JAMA Network Open, vol. 5, no. 6, 2022.

[11] W. H. Lo-Ciganic et al., "Developing and validating a machine-learning algorithm to predict opioid overdose risk," The Lancet Digital Health, vol. 4, 2022.

[12] C. Garbin et al., "Machine learning for predicting opioid use disorder: A review," Computer Methods and Programs in Biomedicine, 2023.

[13] S. L. Song et al., "Predictive models to assess risk of opioid use disorder and overdose: A systematic review," 2024.

[14] K. Nguyen et al., "Design and development of a machine-learning-driven opioid overdose prediction system," 2024.

[15] C. R. Ramírez Medina et al., "A systematic review of machine learning applications in opioid-related harm prediction," npj Digital Medicine, 2025.

[16] Y. S. Liu et al., "Population-level individualized prediction of opioid overdose using machine learning," 2025.