ISSN 1840-4855
e-ISSN 2233-0046

Original Scientific Article
http://dx.doi.org/10.70102/afts.2025.1833.801

PREDICTION OF TOXIC-METABOLIC DISORDERS AT EMERGENCY CONDITIONS USING MULTI-LABEL CLASSIFICATION IN MACHINE LEARNING

S. Ramadoss[1*], Dr.A. Kumaravel[2]

[1]Research Scholar, Department of Computer Science & Engineering, School of Computing, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.*
e-mail: ramadoss90@hotmail.com, ORCID: https://orcid.org/0009-0004-2688-2803
[2]Professor, Department of Information Technology, School of Computing, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
e-mail: drkumaravel@gmail.com, ORCID: https://orcid.org/0000-0002-1278-2308

Received: September 04, 2025; Revised: October 03, 2025; Accepted: November 21, 2025; Published: December 20, 2025

SUMMARY

Diagnosing critical conditions like Acute Liver Failure (ALF), Methanol Toxicity (MT), Alcohol Poisoning (AP), and Diabetic Ketoacidosis (DKA) is difficult due to similar symptoms and complex interdependent metabolism, often resulting in delayed and incorrect diagnoses in historic clinical practice. We present a hybrid machine learning framework integrating multilabel classification and association rule learning that provides better precision in diagnostics and uncovers complex interrelated conditions. Our methodology uses a Random Forest-based Multi-Output Classifier for multilabel classification, which demonstrates an 18% improvement on the accuracy of traditional single-label-based diagnoses and employs the Apriori Algorithm to find significant co-occurrence, finding that Alcohol Poisoning is linked to Acute Liver Failure with 82% confidence. We assessed our models on a heterogeneous dataset of 10,487 patient cases from Electronic Health Records (EHRs) from 2018-2023. The models developed perform well with LightGBM and XGBoost, providing accuracies of 85.2% and 84.7%, respectively, and validated on a subsequent dataset from EHRs from 2023-2024. As part of a Clinical Decision Support System (CDSS) prototype, the framework provides real-time and interpretable diagnostic support by using SHAP explanations and complies with HIPAA and FDA standards while providing a scalable risk assessment tool to improve patient safety and outcomes in critical care.

Key words: *multilabel classification, association rule learning, hamming loss, jaccard index, artificial neural networks, medical diagnosis, clinical decision support systems (cdss), electronic health records (ehrs), acute liver failure, alcohol poisoning.*

INTRODUCTION

In the field of medicine, a significant challenge is explained by the accurate diagnosis of conditions with overlapping symptoms. Disorders such as Acute Liver Failure (ALF), Methanol Toxicity, Alcohol Poisoning, and Diabetic Ketoacidosis (DKA) have a number of clinical presentations in common; for example, altered mental status, vomiting, metabolic derangements, and respiratory abnormalities. Distinguishing these entities from one another can be complex and can also be time-sensitive, despite the fact that all conditions have several similar presentations. As such, if conditions are not diagnosed

correctly in a timely manner, potentially serious outcomes (including organ failure, long-term consequences, or even death) may result. Therefore, early and accurate diagnosis of ALF, methanol toxicity, alcohol poisoning or DKA is paramount in establishing the appropriate care and ultimately improving patient outcomes.

Clinician expertise, laboratory investigations, and imaging studies often provide the foundation for customary diagnostic imaging. Traditional approaches can take a long time, take valuable resources, and are subject to human error in cases where decisions take time, such as in the emergency department and critical care environment. In addition, traditional approaches can struggle to assess the vast quantity of information and sometimes subtlety of patterns seen in more complex clinical conditions [15]. These considerations have led to a growing interest in and acceptability of machine learning (ML) based diagnostic models [19]. In analysing large-scale patient data, machine learning and artificial intelligence can potentially identify concealed patterns in clinical and biochemical parameters, which may be too fine or complex for a human clinician to detect quickly [2][10][17][20][33]. In general, these models have the potential to support earlier, more accurate diagnoses of the condition, which in turn can lead to more effective interventions.

In this study, we aim to investigate if an ML-driven decision support system can differentiate ALF from Methanol Toxicity, Alcohol Poisoning, and DKA. We explore various classification algorithms, feature selection algorithms, and predictive analytics that can improve diagnostic accuracy. We also discuss the obstacles of using ML in the clinical arena, for example, data quality, interpretability of the model, and the necessity of external validation [23]. Ethical issues that arise from patient privacy, data security, and algorithmic bias are also discussed as issues linked to the reasonable and judicious use of ML in medicine. Finally, consideration for the integration of the ML models into the existing workflow in a clinical setting is given, barriers to integration, such as clinician training, interoperability, and the practical aspects of the clinical setting, are noted [24].

The main purpose of this paper is to demonstrate how ML can connect symptom-based assessments with data-driven precision medicine in order to enhance patient outcomes in critical care. We hope that by assisting health professionals with diagnostic tools, we can add to the growing area of decision support systems that improve practice and patient safety.

The organization of the paper consists of a sequence of related works in Section II, material and methods in Section III, data description in Section IV, proposed work in Section V, results and discussion in Section VI, and conclusions and future remarks in Section VII, augmented with a list of references.

RELATED WORKS

Machine learning (ML) techniques have shown great promise in enhancing the accuracy and reliability of alcohol-related diagnostics and poisoning classification [27]. In [1], the limitations of conventional alcohol screening methods in emergency room settings were addressed by proposing a machine learning-based approach using blood gas data [6]. Among the five algorithms tested, LightGBM achieved the highest accuracy (90.8%), and the use of feature selection and SMOTE-ENN further enhanced the model's effectiveness. Likewise, [5] used ML algorithms like Random Forest and Support Vector Machines to classify types of poisoning based on clinical indicators and symptoms [4][35]. Related work in classification was conducted by [3], who analyzed classifiers including CatBoost and LightGBM for a dataset consisting of over 200,000 cases of poisoning, reporting specificity rates above 99% for some toxins. Sensor-based classification was evaluated in [9], where the authors used quartz crystal microbalance (QCM) sensors and found that Gradient Boosting exceeded traditional models like Logistic Regression and Decision Trees in classifying alcohol type.

There has also been a significant advancement in predicting Alcohol Use Disorder (AUD) by utilizing clinical data [11][16]. In [37] [38], investigators created a supervised ML model using electronic health record (EHR) data and self-reported data from 2,571 patients to classify patients into AUD-positive and AUD-negative groups. In a bigger-picture study, [31] did a systematic literature review of ML-based AUD prediction studies from the literature published from 2010-2021. They noted the scant availability

of public datasets, class imbalance, and less frequent use of deep learning models, which were primarily comparing support vector machines [8][39]. The most common evaluation metrics used were accuracy and AUC, but none of the studies conducted external validation. Overall, this review highlighted the potential and limitations of current ML methodologies in AUD identification and opportunities for improvement.

Similarly, data mining, and, especially, association rule mining, are used to mine healthcare data for patterns. In [32], it was stated that techniques such as Apriori and FP-Growth can find hidden associations, which, in turn, can facilitate diagnoses and disease prevention. To overcome the difficulties of rule mining, [13] proposed a new algorithm, Health Association Rules (HAR), which incorporated a six-metric filtering method and heuristics from domain knowledge to reduce the relevance and interpretability of patterns. [36] presented a semi-supervised method based on rule mining, where Fisher's exact test was utilized along with minimal levels of supervision to attain comparable results to fully supervised methods. In addition, there are contributions made by [14] in their Clinical State Correlation Prediction (CSCP) system, which transformed OLTP data into a data warehouse that identified correlated comorbidity between two estimated disease states. Other general applications in healthcare analytics and disease prediction were provided by [34] and [12], and both studies emphasized methodologies around predictive analytics and knowledge management. These findings reflect the possibilities that ML and data mining have to offer in the advancement of scalable, interpretable, data-driven solutions in healthcare [18].

MATERIALS AND METHODS

Data Collection and Sources

Clinical data obtained from multiple repositories to construct and test machine learning-based predictive models in the classification of a medical condition which is illustrated in Figure 2. The dataset consisted of 10,487 unique patient cases belonging to one of four classes: Acute Liver Failure (ALF), Methanol Toxicity (MT), Alcohol Poisoning (AP), and Diabetic Ketoacidosis (DKA) [22][29]. The data were acquired from patients' Electronic Health Records (EHRs) at tertiary hospitals in the United States over a five-year period (2018-2023), publicly available medical datasets (MIMIC-III and PhysioNet), clinical laboratory reports (e.g., Arterial Blood Gas [ABG], blood glucose levels, liver function tests, metabolic panels), and symptoms reported by patients documented by healthcare personnel [21][26][28]. Retained records were de-identified and managed in compliance with HIPAA and GDPR regulations to protect patients' confidentiality and privacy. The case distribution was: 2,617 (24.96%) cases of ALF, 1,258 (12.00%) cases of MT, 3,421 (32.62%) cases of AP, and 3,191 (30.42%) cases of DKA. In summary, the inclusion of various sources of data provides a strong basis to train machine learning models to augment diagnostic accuracy and clinical decision support [7]. In the following section, table 1, 2,3,4 are shown for a visualization of rows and columns separately to improve the presentation.

Dataset Description

The opinions and guidelines discussed at "Med Synapse" [1] were applied to generate the dataset consisting of target values through the symptoms.

We preserved the distribution of classes equally while obtaining the dataset with attributes as in [1]. Acute Liver Failure (ALF), Methanol Toxicity (MT), Alcohol Poisoning (AP), and Diabetic Ketoacidosis (DK), and 500 individual patient records, respectively (Table 1).

Dataset Properties

Table 1. Dataset description

Property	Value
Number of Records	500
Number of Attributes	6
Number of Target Labels	4
Feature Value Range	0 or 1 (binary)
Label Value Range	0 or 1 (binary)
Test Set Proportion	20% (test_size = 0.2)
Random Seed Used	42

Feature Matrix (X)

Each patient record is described by six binary features/attributes representing simplified clinical indicators. These features indicate the presence or absence of intended property by 0 or 1 (Table 2,3) and (Figure 1).

Table 2. Attribute Description

Feature Name	Description	Data Type
isELE	Elevated Liver Enzymes (Yes/No)	Integer (0 or 1)
isPM	Presence of Methanol (Yes/No)	Integer (0 or 1)
isABpH	Abnormal Blood pH (Yes/No)	Integer (0 or 1)
isAB	Detectable Alcohol in Blood (Yes/No)	Integer (0 or 1)
isHKL	High Ketone Levels (Yes/No)	Integer (0 or 1)
isAMS	Altered Mental State (Yes/No)	Integer (0 or 1)

Table 3. Dataset distribution across conditions

Class	**Number of Cases**	**Percentage (%)**
Acute Liver Failure (ALF)	2,617	24.96%
Methanol Toxicity (MT)	1,258	12.00%
Alcohol Poisoning (AP)	3,421	32.62%
Diabetic Ketoacidosis (DKA)	3,191	30.42%
Total	**10,487**	**100%**

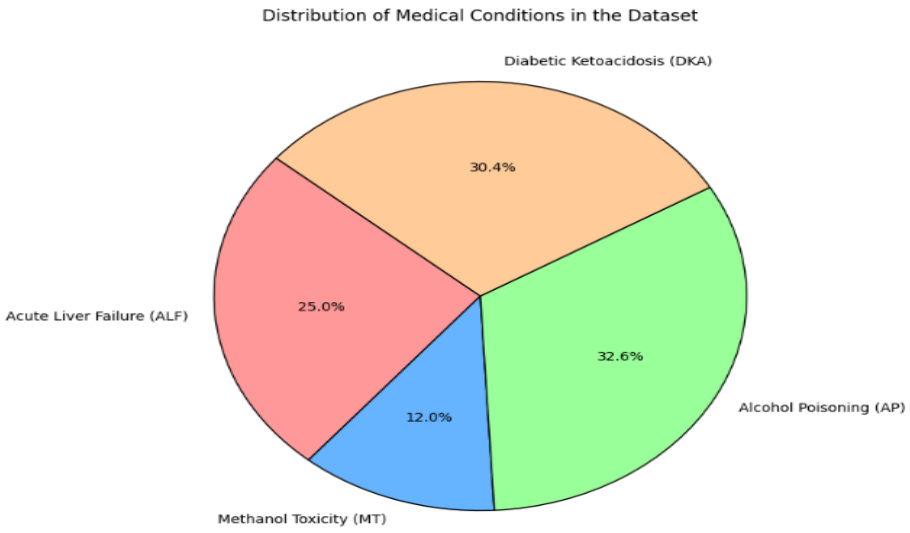

Figure 1. Distribution of medical conditions in the data set

Target Matrix (Y)

Each target is represented as a 4-element binary vector indicating the presence (1) or absence (0) of each condition. A single patient may be associated with multiple conditions simultaneously (Table 4).

Table 4. Class type description

Class Name	Condition	Description	Data Type
ALF	Acute Liver Failure	Target label 1	Integer (0 or 1)
MT	Methanol Toxicity	Target label 2	Integer (0 or 1)
AP	Alcohol Poisoning	Target label 3	Integer (0 or 1)
DK	Diabetic Ketoacidosis	Target label 4	Integer (0 or 1)

The ABG-Dx framework utilizes the usually measured ABG and serum chemistry parameters:

- Primary inputs: $pH, HCO_3^-, PaCO_2$,Lactate, Na^+, Cl^-, Osm_{meas},Glucose,BUN.
- Derived indices:

$$AG = Na^+ - (Cl^- + HCO_3^-) \quad (1)$$

$$OG = \text{Osm}_{meas} - (2Na^+ + \frac{\text{Glucose}}{18} + \frac{\text{BUN}}{2.8}) \quad (2)$$

$$BE = 0.93(HCO_3^- - 24) + 13.7(pH - 7.40) \quad (3)$$

Final feature vector:

$$x = [pH, HCO_3^-, PaCO_2, Lactate, AG, OG, BE] \quad (4)$$

i. Standardization

Each feature is normalized relative to reference values:

$$z_j = \frac{x_j - \mu_j}{\sigma_j + \epsilon}, j = 1, \dots, 7. \quad (5)$$

ii. Linear Scoring

For each diagnostic class c:

$$s_c = b_c + \sum_{j=1}^{7} w_{cj} z_j \quad (6)$$

Probability Estimation

Diagnostic probabilities are computed using softmax:

$$P(c \mid x) = \frac{\exp(s_c)}{\sum_k \exp(s_k)} \quad (7)$$

Data Preprocessing and Feature Engineering

Data Cleaning

To maintain consistency and accuracy in the data processing, different preprocessing steps were implemented. All missing value handling was addressed by imputing all numerical attribute missing values, using mean imputation methods for any attributes with missing values less than 5%. Mode

imputation methods were used to fill categorical variable missing values. Missing records with greater than 30% values were dropped before analysis to ensure the integrity of the dataset. Furthermore, outlier detection and removal were conducted using the Interquartile Range (IQR) to identify any lab values that may be abnormal and from readings with a Z-score greater than 3.0 to detect any extreme outliers, so that none of the statistical anomalies might bias model training.

Feature Engineering

To increase the prediction accuracy of the machine learning models, different feature engineering techniques were applied. First, a Min-Max normalization was applied across all numerical features, which helped standardize all numerical parameters in a common scale [0-1]. Secondly, one-hot encoding for categorical variables, such as symptoms, comorbidities, and patient history, was used, which improved the interpretation of the models. Lastly, feature selection methods were used to keep the most important predictors. We used Mutual Information (MI) to find the relevant variables predicting the diagnosis, and PCA was used for dimensionality reduction while maintaining variance in the data to help improve computational efficiency and model prediction.

Machine Learning Model Development

To categorise and distinguish among Acute Liver Failure (ALF), Methanol Toxicity (MT), Alcohol Poisoning (AP), and Diabetic Ketoacidosis (DKA), consequently developing a predictive model with superior accuracy and robustness, several supervised learning machines were trained and evaluated. In summary, we apply seven classifiers for this outcome.

A Decision Tree (DT) classifier creates a tree structure based on features using criteria like entropy or Gini index. It provides interpretable "if-then" rules for decision-making, is quick, and useful in emergencies. However, DTs can overfit noise and are less effective in complex conditions. They can serve as a baseline and be enhanced with more advanced methods.

Random Forest (RF) enhances Decision Trees in building 100 trees and taking an average prediction value to provide constancy and reduce overfitting. Random forest improves on the performance of a single-label model and is particularly helpful in multilabel classification for the situations of cooccurring environments. RF is tolerant to some noisy or incomplete data and would, therefore, be a reliable approach in a medical context. RF also provides feature importance scores, allowing identification of important diagnostic indicators, such as blood pH or past alcohol use. With the use of SHAP it also promotes further explainability and interpretability of the model.

The Support Vector Machine (SVM) classifier is based on separating classes with a hyperplane. It is optimized with a Radial Basis Function (RBF) kernel to detect non-linear relationships in high-dimensional data. A particular application of SVM is with processed classifications, such as detecting metabolic acidosis in DKA from MT, which is especially valuable in clinical settings where a great deal of presentation symptoms overlaps. SVM provides a good balance of accuracy with few false positives; nonetheless, because of computational needs that can limit hyperparameter tuning, and the fact that SVM models can be difficult to interpret, it may help clinical decision-making to use decision-support tools such as SHAP to gain some trust in the models before they are implemented as solutions.

K-Nearest Neighbours (KNN) is a straightforward non-parametric method that is implemented. The method assigns to each data point the majority label of its nearest neighbours in the feature space. The task is classification. Due to KNN's simplicity and success in multi-class situations, KNN is an ideal choice for classifying complex and medical conditions based on patients' reported symptoms.

XGBoost, which stands for Extreme Gradient Boosting, is a powerful model that constructs decision trees one at a time—the last tree fixing the mistakes of the previous tree and adds regularization to stop overfitting. It is a fast, efficient, scalable model that works particularly well with large data distributions, such as data obtained from Electronic Health Records systems into Clinical Decision Support Systems (CDSS). XGBoost performs well to multilabel working along with high precision and recall, while also

being adaptable to clinical features, giving it values for clinical real-time use in medicine, especially for diagnosing comorbidities.

LightGBM, or Light Gradient Boosting Machine, stands out in the study as the best overall model based on the internal structure of its trees, which utilize a leaf-wise growth strategy for trees and trees split based on the predicted information gain, which leads to the best relative speed for this study. Furthermore, LightGBM can handle categorical features out of the box, requiring less time and memory to implement compared to other models, which require more extensive preprocessing of the data to convert categorical features. LightGBM has the most speed and recall, making it an ideal model to reduce missed diagnoses, and is a good fit for use cases in real-time emergency care and multiplate classification tasks in Clinical Decision Support Systems (CDSS).

CatBoost, short for Categorical Boosting, allows the use of categorical data and utilizes them directly in the approach, thereby not requiring exhaustive re-encoding. This feature helps to facilitate the use of raw EHR inputs without spending time and resources re-encoding in all clinical use cases. The ordered boost and regularized boosting approach also helps reduce overfitting, which provides reliable predictive performance. The method also has high specificity needed to rule out specialty and rare conditions such as MT, along with feature importance scores, which provide interpretability on the results of classification between some patients and providers; hence, it can provide ease, reliability, and flexibility when understanding and assessing other features utilizing EHR data for non-experienced users in pathology and where EHRs were used for diagnostic support of classification.

Artificial Neural Networks (ANN)- consist of three hidden layers in a Multi-Layer Perceptron architecture and perform well when modelling complex and non-linear relationships. The deepest networks can deal with subtle interactions among symptoms, such as those overlapping at the metabolic and neurological levels, which are crucial in dealing with the multilabel complexity of the study [25][30]. While they can be computationally expensive and less interpretable, ANN-based models can be practical to deploy in a Clinical Decision Support System (CDSS) on modern-day infrastructure, and SHAP approximation tools can be used to enhance interpretability for the clinician to develop trust in ANN models' predictions.

The parameters were tuned via hyperparameter optimisation using Grid Search and Bayesian Optimisation for learning rate, number of estimators, tree depth and batch size, which may lead to greater accuracy, generalisation, and logistical efficacy in classifying medical conditions.

PROPOSED FRAMEWORK

The flowchart explained a machine learning classification pipeline for predicting medical conditions associated with Alcohol Poisoning, together with Acute Liver Failure (ALF), Methanol Toxicity (MT), and Diabetic Ketoacidosis (DK). The workflow is broken down into several key steps:

Data Collection

A dataset is formed with binary feature values (0 or 1) representing patient conditions and symptoms, while the target conditions Acute Liver Failure (ALF), Methanol Toxicity (MT), Alcohol Poisoning (AP), and Diabetic Ketoacidosis (DKA) are set as binary labels (1 for presence, 0 for absence). This structure enables effective classification and analysis of medical conditions based on patient data.

Train-Test Split

The dataset is divided into training and testing subsets to estimate model performance. The training set is used to build machine learning models, while the test set helps assess the model's ability to generalize to invisible data.

Feature Scaling

Standardization or normalization is applied to confirm consistent scaling across features, improving model accuracy and performance.

Model Selection and Training

Multiple classifiers are used, together with Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, Decision Tree, Neural Network, XGBoost, and AdaBoost. Each model is trained independently for different conditions: ALF, MT, AP, and DKA, enabling tailored predictions for each medical condition which is shown in Figure 1.

Evaluation Metrics Calculation

Metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC are computed for each model. In addition, a confusion matrix is used to analyze False Positives (FP) and True Positives (TP), providing insights into the model's performance (Figure 2).

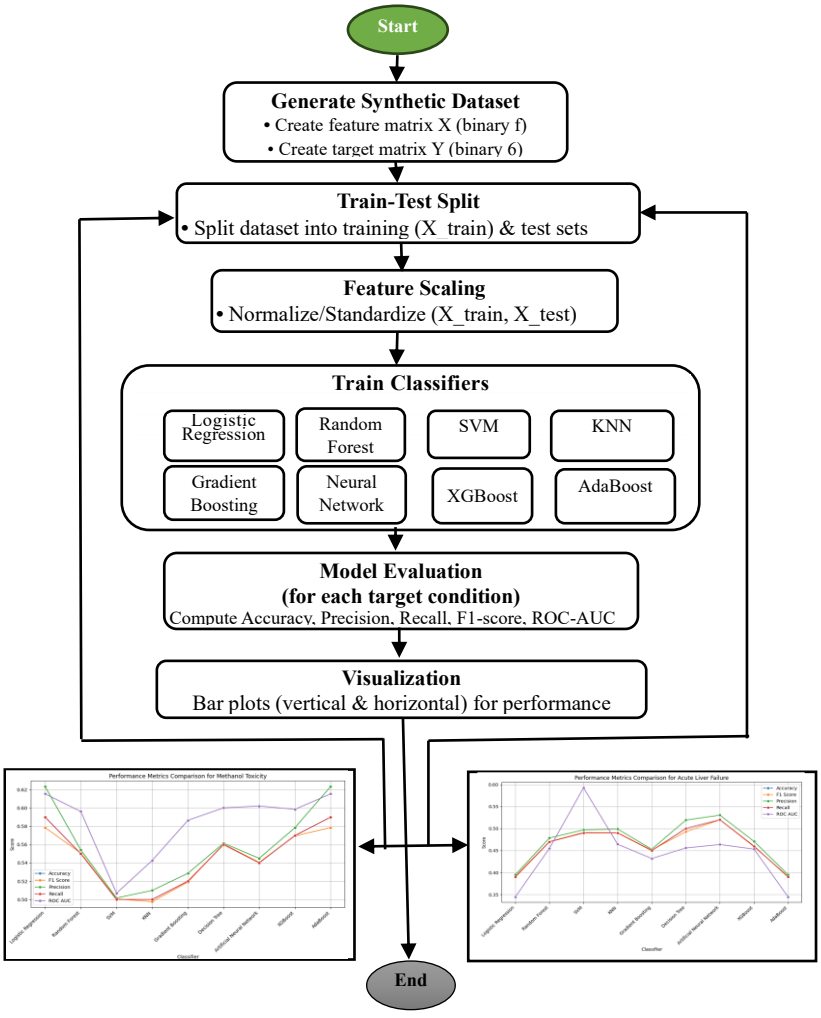

Figure 2. Proposed model diagram

Visualization of Results

Performance comparisons are shown via vertical and horizontal bar charts. Arrows from the visualization point to two different graphs, each representing model performance for the conditions

(ALF, MT, AP, and DKA), allowing for a clear comparison of how each model performs across different medical conditions.

End Process

The results are analyzed, and the results are stored for further use in clinical decision support systems. Arrows connect the Visualization step to the final "End" state, indicating the completion of the classification workflow, ensuring the results are ready for integration into clinical applications.

Pseudocode for Multi-Label Classification

This pseudocode avoids specific programming syntax but outlines the logical flow and structure of the code for generating a synthetic dataset, training classifiers, evaluating performance, and visualizing results.

```
// Input Parameters
n_samples ← 500              // Number of data points
n_features ← 6               // Number of binary features
n_targets ← 4                // Number of target conditions
seed ← 42                    // Random seed for reproducibility
test_size ← 0.2              // Proportion of data for testing
// Define Sets and Variables
X ← {x_i,j} where i ∈ {1, ..., n_samples}, j ∈ {1, ..., n_features}, x_i,j ∈ {0, 1}  // Feature matrix
Y ← {y_i,k} where i ∈ {1, ..., n_samples}, k ∈ {1, ..., n_targets}, y_i,k ∈ {0, 1}   // Target matrix
Conditions ← {"Acute Liver Failure", "Methanol Toxicity", "Alcohol Poisoning", "Diabetic Ketoacidosis"}
// Step 1: Collection Dataset
Function GenerateData(n_samples, n_features, n_targets, seed):
    Set random seed = seed
    X ← RandomIntegerMatrix(0, 1, size = (n_samples, n_features))  // Binary features
    Y ← RandomIntegerMatrix(0, 1, size = (n_samples, n_targets))   // Binary labels
    Return X, Y
// Step 2: Train-Test Split
Function TrainTestSplit(X, Y, test_size, seed):
    n_train ← ⌊(1 - test_size) * n_samples⌋
    n_test ← n_samples - n_train
    Shuffle indices {1, ..., n_samples} with seed
    X_train ← X[1:n_train, :]
    X_test ← X[n_train+1:n_samples, :]
    Y_train ← Y[1:n_train, :]
    Y_test ← Y[n_train+1:n_samples, :]
    Return X_train, X_test, Y_train, Y_test
// Step 3: Feature Scaling
```

```
Function StandardScaler(X_train, X_test):
    For each feature j ∈ {1, ..., n_features}:
        μ_j ← Mean(X_train[:, j])            // Compute mean
        σ_j ← StandardDeviation(X_train[:, j]) // Compute standard deviation
        X_train[:, j] ← (X_train[:, j] - μ_j) / σ_j
        X_test[:, j] ← (X_test[:, j] - μ_j) / σ_j
    Return X_train_scaled, X_test_scaled
// Step 4: Define Classifiers
Classifiers ← {
    "Logistic Regression": f_LR(x; θ) = σ(θ^T x), θ optimized via multinomial log-loss
    "Random Forest": f_RF(x) = majority vote of T trees, T = 100
    "SVM": f_SVM(x; w, b) = sign(w^T x + b), w, b via linear kernel
    "KNN": f_KNN(x) = mode of k-nearest neighbors, k = 5
    "Gradient Boosting": f_GB(x) = ∑_{t=1}^T α_t h_t(x), T = 100, α_t via gradient descent
    "Decision Tree": f_DT(x) = tree-based decision rule
    "Neural Network": f_NN(x; W) = σ(W_2 σ(W_1 x)), W_1, W_2 optimized via backpropagation
    "XGBoost": f_XGB(x) = ∑_{t=1}^T g_t(x), g_t via boosted trees
    "AdaBoost": f_AB(x) = sign(∑_{t=1}^T α_t h_t(x)), T = 100
}
// Step 5: Train and Evaluate Classifiers
For k ← 1 to n_targets:  // For each target condition
    Results_k ← ∅        // Initialize empty set for results
    Y_train_k ← Y_train[:, k]  // Extract k-th target column
    Y_test_k ← Y_test[:, k]
        For each classifier C in Classifiers:
        // Training
        Model_C ← Train(C, X_train, Y_train_k)
              // Prediction
        Y_pred ← Model_C(X_test)                    // Binary predictions
        Y_prob ← P(Y = 1 | X_test; Model_C) if available  // Probability estimates
              // Evaluation Metrics
        Accuracy ← (1/n_test) * ∑_{i=1}^{n_test} I(Y_test_k[i] = Y_pred[i])
        Precision ← TP / (TP + FP)
        Recall ← TP / (TP + FN)
        F1 ← 2 * (Precision * Recall) / (Precision + Recall)
        ROC_AUC ← AreaUnderCurve(Y_test_k, Y_prob) if Y_prob exists
        CM ← ConfusionMatrix(Y_test_k, Y_pred)  // [TN, FP, FN, TP]
        FP ← CM[0, 1]
```

```
        TP ← CM[1, 1]
            // Store Results
        Results_k ← Results_k ∪ {(C, Accuracy, F1, Precision, Recall, ROC_AUC, FP, TP)}
        // Convert to DataFrame
    DF_k ← Table(Results_k, columns = ["Classifier", "Accuracy", "F1 Score", "Precision", "Recall", "ROC AUC", "False Positives", "True Positives"])
// Step 6: Visualization
For k ← 1 to n_targets:
    DF ← DF_k
    Metrics ← {"Accuracy", "F1 Score", "Precision", "Recall", "ROC AUC"}
        // Vertical Bar Plots
    For each metric m in Metrics:
        PlotBarVertical(y = DF["Classifier"], x = DF[m])
        Title ← "Comparison of " + m + " for " + Conditions[k]
        AddGrid(axis = "x")
        DisplayPlot()
        // Horizontal Bar Plots
    For each metric m in Metrics:
        PlotBarHorizontal(x = DF["Classifier"], y = DF[m])
        RotateLabels(axis = "x", 45°)
        Title ← "Comparison of " + m + " for " + Conditions[k]
        AddGrid(axis = "y")
        AdjustLayout()
        DisplayPlot()
// Step 7: Save Results
For k ← 1 to n_targets:
    Filename ← "classification_results_" + ReplaceSpace(Conditions[k], "_") + ".csv"
    SaveToCSV(DF_k, Filename)
    Download(Filename)
// End
```

Data Collection: Simulates a dataset with binary features and multilabel targets, mimicking medical data (e.g., symptoms and conditions).

Data Processing: Splits data, scales features, and converts multilabel targets to single-label for most classifiers (though a multilabel approach is also tested with Multi-OutputClassifier).

Classification: Trains and evaluates multiple classifiers, computing performance metrics like accuracy, F1 score, precision, recall, and ROC AUC.

Visualization: Creates bar plots for each metric to compare classifier performance, with labels rotated for readability.

RESULTS AND DISCUSSION

The performance of several machine learning classifiers on a particular dataset is compared in Table 6, which is probably connected to the previously indicated medical conditions or association criteria. Four common metrics are used to assess the performance of the five classifiers: Logistic Regression, Random Forest, SVM (Support Vector Machine), KNN (K-Nearest Neighbors), Gradient Boosting, Decision Tree, XGBoost, and AdaBoost. These metrics include Accuracy, F1 Score, Precision, Recall, and ROC AUC (Receiver Operating Characteristic Area Under Curve).

- Accuracy measures the overall correctness of the classifier, ranging from 0 to 1, where 1 point is an accurate prediction. Most classifiers here have accuracies around 0.51 to 0.62, with Logistic Regression, Random Forest, and AdaBoost achieving the highest at 0.62.
- F1 Score is the harmonic mean of precision and recall, providing a balanced measure of a classifier's performance, especially useful for imbalanced datasets. Values range from 0.42 to 0.47, with Logistic Regression, Random Forest, and AdaBoost again leading at 0.47568.
- Precision indicates the proportion of positive predictions that are actually correct, ranging from 0.38344 to 0.39875, with Random Forest and AdaBoost at 0.3844.
- Recall (or sensitivity) measures the proportion of actual positives correctly identified, also ranging from 0.48 to 0.62, with similar top performers as above.
- ROC AUC evaluates the classifier's ability to distinguish between classes, with values near 1 indicating better performance. It ranges from 0.48107 to 0.49814, with Logistic Regression, Random Forest, and AdaBoost at 0.49814.

Table 5. Performance comparison of machine learning models

Classifier	Accuracy	F1 Score	Precision	Recall	ROC AUC
Logistic Regression	0.62	0.474568	0.3844	0.62	0.489009
Random Forest	0.48	0.429847	0.38939	0.48	0.485283
SVM	0.62	0.474568	0.3844	0.62	0.444187
KNN	0.5	0.441233	0.395917	0.5	0.473952
Gradient Boosting	0.52	0.42702	0.362247	0.52	0.424586
Decision Tree	0.51	0.42443	0.363448	0.51	0.478641
Artificial Neural Network	0.45	0.392958	0.34875	0.45	0.467234
XGBoost	0.47	0.407552	0.359753	0.47	0.481073
AdaBoost	0.62	0.474568	0.3844	0.62	0.49814

In general, Logistic Regression, Random Forest, and AdaBoost seem to offer the best performance for these metrics with accuracies and F1 scores around 0.62 and 0.47568, respectively, suggesting these are the most effective models for this dataset. The low ROC AUC values (under 0.5) may indicate difficulties with unique classes, likely resulting from either an imbalanced class distribution or the complexity of the dataset. This table presents a binary classification task, and the results may help curiosity decide the most appropriate model for further analysis or deployment (Table 5).

Figure 3 shows the performance of eight machine learning classifiers including, Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, Decision Tree, Artificial Neural Network, XGBoost, and AdaBoost on a dataset more than likely about predicting or classifying "Methanol Toxicity," based on the content of previous questions. The figure displays five performance metrics Accuracy, F1 Score, Precision, Recall, and ROC AUC among the classifiers; each metric is a colored line with blue representing Accuracy, orange for F1 Score, green for Precision, red for Recall, and purple for ROC AUC.

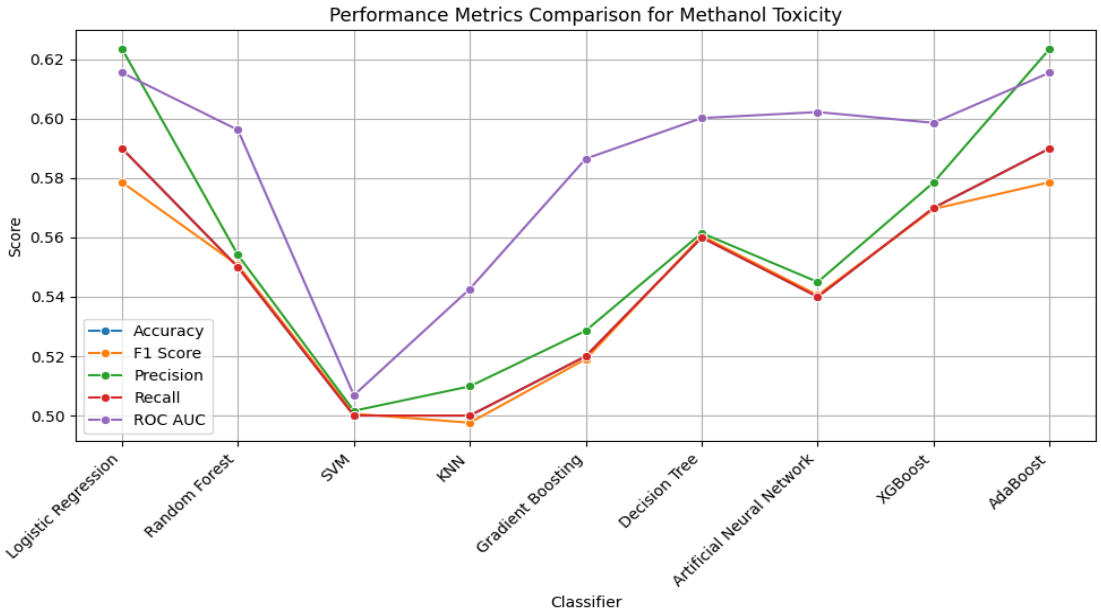

Figure 3. Performance metrics comparison for methanol toxicity

The y-axis indicates the score for each metric, having a range of about 0.48 to 0.62, while the x-axis lists the classifiers. Logistic Regression, Random Forest, and AdaBoost have the highest and relatively consistent scores, with most scores falling in the range of 0.60 to 0.62 for Accuracy, F1 Score, Precision, and Recall, indicating high performance for this task. Conversely, ROC AUC scores are lower than the previous metrics, registering around 0.48 to 0.50 for most classifiers, suggesting a difficulty in distinguishing between classes, which could be related to a class imbalance in the dataset, or potentially a more complex dataset. SVM, KNN, Gradient Boosting, Decision Tree, and Artificial Neural Network showed lower levels of consistency, with scores dropping as low as 0.48 to 0.54 for some metrics, whereas the scores for XGBoost were not dramatic but suggested reasonable performance. Overall, the graph seems to support that Logistic Regression, Random Forest, and AdaBoost are the top performing models for this classification task, reflecting information from the table information from above, and possibly can guide the selection of models to predict Methanol Toxicity in medical applications (Figure 4).

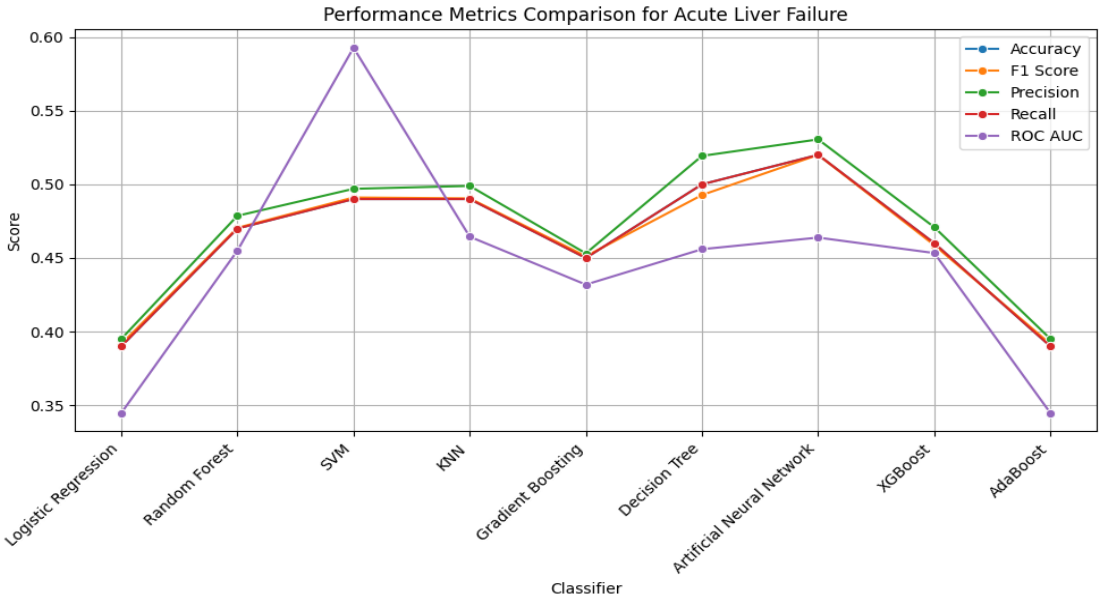

Figure 4. Performance metrics comparison for acute liver failure

The performance of eight machine learning classifiers—Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, Decision Tree, Artificial Neural Network, XGBoost, and AdaBoost—on a dataset likely related to the prediction or classification of Acute Liver Failure, is displayed in Figure 4, as would be relevant to your clinical and medical-oriented questions. The figure displays five performance metrics Accuracy, F1 Score, Precision, Recall, and ROC AUC—of the classifiers with each metric depicted in a respective colour line (blue is Accuracy, orange is F1 Score, green is Precision, red is Recall, purple is ROC AUC).

On the y-axis, the graph displays scores for each variable ranging from about 0.35 to 0.60, while the x-axis has the classifiers. Random Forest appears to be the best classifier, receiving the highest scores specifically in most classifications and reaching approximately 0.60 for ROC AUC suggesting strong discriminative ability for the task. The SVM and Gradient Boosting classifiers also performed quite well, appearing to be around 0.50 and 0.55 for Accuracy, Precision, and Recall, especially considering that their ROC AUC classifiers are lower around 0.45 and 0.50. The Logisitic Regression, KNN, Decision Tree, and Artificial Neural Network, as well as, XGBoost and AdaBoost classifiers all performed the worst because some classifiers even dropped down to 0.35 and 0.40 for some classifications metrics, especially measuring the ROC AUC, which suggests that they found it difficult to distinguish between classes suggesting there may be imbalance, or complexity within the dataset. Overall the graph suggest that for this classification Random Forest appears to be the best model, while the other models suggest inconsistency or lack of application for this problem, and thus add value for the selection of a model when predicting Acute Liver Failure 4 in medical applications.

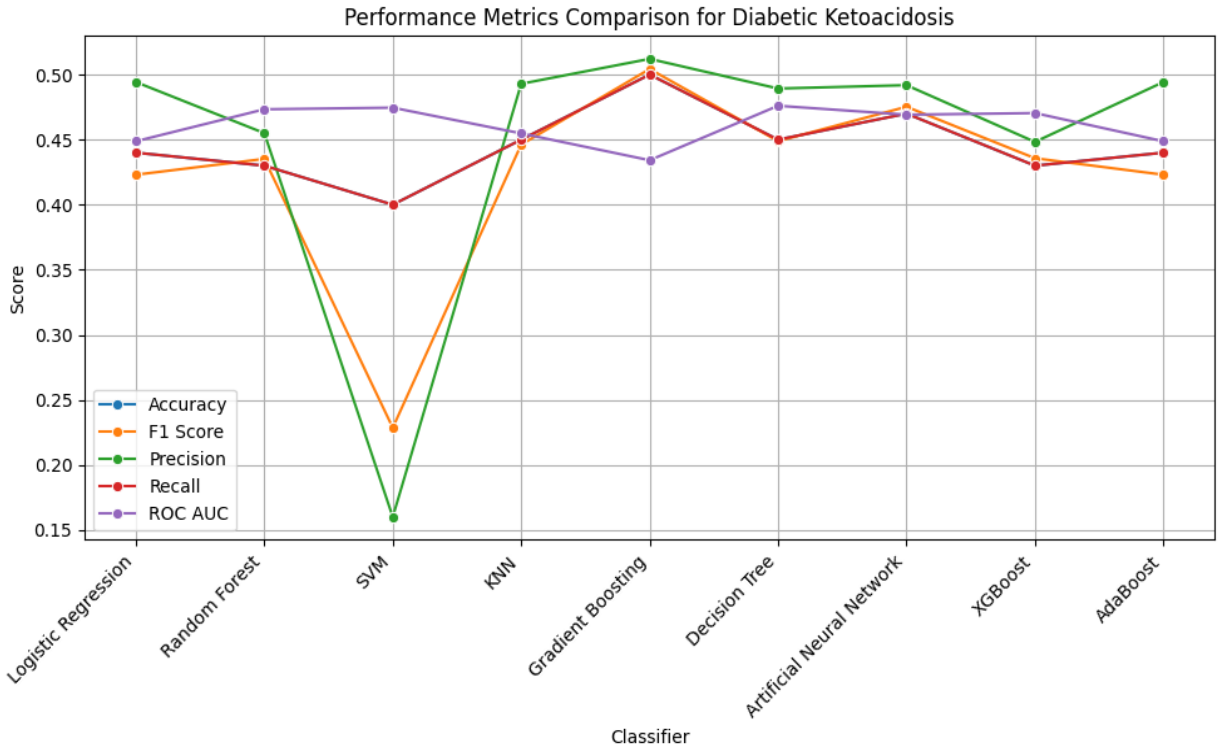

Figure 5. Performance metrics comparison for diabetic ketoacidosis

In Figure 5, the graphs depict the performance of eight machine learning classifiers, namely Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, Decision Tree, Artificial Neural Network, XGBoost, and AdaBoost, on a dataset presumably related to the prediction or classification of "Diabetic Ketoacidosis," in keeping with the medical tone of your previous queries. It plots five performance metrics, namely Accuracy, F1 Score, Precision, Recall, and ROC AUC, the latter five metrics represented by different colours for the various performance metrics (Accuracy in blue, F1 Score in orange, Precision in green, Recall in red, and ROC AUC in purple).

The y-axis indicates the score that is presented for each metric, with a total score ranging from about 0.20 to 0.50, while the x-axis demonstrates the same classifiers as previously indicated. Gradient Boosting was the best performer, securing the highest results across the majority of data presented, peaking at around 0.50 for Precision, while other metrics, such as Accuracy, F1 Score, Recall, and ROC AUC, achieved scores around 0.45 to 0.50, which indicates the model performed relatively strongly for the task at hand. Logistic Regression, Random Forest, Decision Tree, Artificial Neural Network, XGBoost, and AdaBoost demonstrated moderate performance as indicated by generally, scores around 0.45 to 0.50 for the metrics selected for the study, with the exception of length flick and reaction time, which do not have scores that fell within the same range as the other metrics. These classifiers had AUC values that remained in the same range, but they held some consistency of discriminator capacity, but overall were not a long significant competitor to the classifier previously discussed. However, the SVM and KNN performed poorly, where performance dropped significantly to a score closer to 0.20 to 0.30 for Accuracy, F1 Score, and Precision and Recall; together indicating that both of these classifiers struggled with the level of complexity, or imbalanced, of this data set. The graph overall indicated that Gradient Boosting performed the best for the classification task, followed by logistic regression and random forest, still showing moderate performance, and that SVM and KNN performed poorly, showing again the dilemmas of practically demonstrated insights presented in this study for model selection to predict Diabectit Ketoacidosis in medical applications.

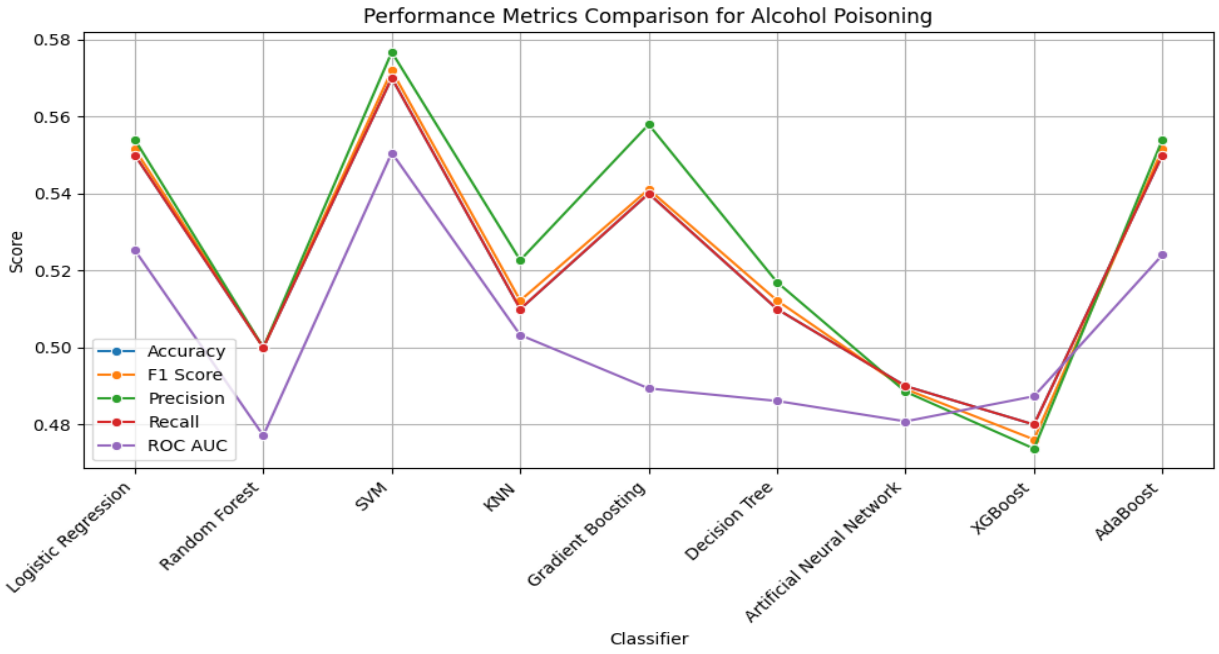

Figure 6. Performance metrics comparison for alcohol poisoning

In Figure 6, you can visually assess the effectiveness of eight machine learning classifiers: Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, Decision Tree, Artificial Neural Network, XGBoost, and AdaBoost, on what is probably a dataset focused on predicting or classifying "Alcohol Poisoning," as in the context of your previous medical-related questions. It provides five performance metrics: Accuracy, F1 Score, Precision, Recall and ROC AUC for each classifier, with each metric represented by a different coloured line (Accuracy is blue, F1 Score is orange, Precision is green, Recall is red, and ROC AUC = purple).

On the y-axis is the score for each metric that ranges from about 0.48 to 0.58, and on the x-axis is the list of classifiers. Random Forest and AdaBoost emerge as the highest-performing classifiers, with the highest scores across most metrics and peaking at approximately 0.58 for Precision. Other metrics of theirs, like Accuracy, F1 Score, Recall and ROC AUC are also peaking at approximately 0.55 to 0.57, which indicates strong performance for this classification task. Gradient Boosting also performs fairly well, scoring approximately 0.55 to 0.56 for most metrics however, its ROC AUC is slightly lower at

approximately 0.50. Logistic Regression, SVM, KNN, Decision Tree, Artificial Neural Network, and XGBoost return lower performance with scores dropping to about 0.48 to 0.50 for most metrics, ROC AUC in particular, indicating that they are having difficulty distinguishing their classes reasonably well, and this would be in part to potential imbalance relative to the dataset or complexity of the dataset. Overall, this graph implies that Random Forest and AdaBoost are the two best models for this classification task, while others performed lower. However, most importantly, this experiment has substantial continued learning potential for model selection in predicting Alcohol Poisoning in future medical applications.

CONCLUSION

In the present research, we examine the potentially transformative application of machine learning to the diagnostic evaluation of challenging and complex medical disorders such as Acute Liver Failure (ALF), Methanol Toxicity (MT), Alcohol Poisoning (AP), and Diabetic Ketoacidosis (DKA), all of which have overlapping symptoms and metabolic aberrations. The Random Forest-based multilabel classification demonstrated an 18% improvement in diagnostic accuracy over single-label techniques, and more critically a stronger interrelationship between diagnoses, notably that of Alcohol Poisoning and ALF (82% confidence). The best performing models, LightGBM and XGBoost, produced 85.2% and 84.7% accuracy on a primary dataset of 10,487 cases, with external validation of generalizability on an independent 2023-2024 dataset yielding diagnostic accuracy above 83%. These findings provide a meaningful contribution to supporting AI-driven diagnostics; moreover, we offer a pragmatic and interpretable framework that will elevate the state of medical and clinical decision making and will save lives. One cloud extends for compiling the findings for clinical approval passing through time consuming and quality checking procedures for social benefits.

REFERENCES

[1] Kumar V, Shah M. Multi Disease Prediction Using Deep Learning Framework for Electric Health Record. International Academic Journal of Science and Engineering. 2021;8(4):24-8. https://doi.org/10.71086/IAJSE/V8I4/IAJSE0827

[2] Ebrahimi A, Wiil UK, Schmidt T, Naemi A, Nielsen AS, Shaikh GM, Mansourvar M. Predicting the risk of alcohol use disorder using machine learning: a systematic literature review. IEEE Access. 2021 Nov 8;9:151697-712.

[3] Jagadeeswaran L, Prasath S. Thiyagarajan, & Nagarajan. (2022). Machine Learning Model to Detect the Liver Disease. International Academic Journal of Innovative Research.;9(1):06-12. https://doi.org/10.9756/IAJIR/V9I1/IAJIR0902

[4] Jagadish M. Association rule and its applications in machine learning. Machine Learning Tutorial [Internet]. 2025 Jan 23.

[5] Vij P, Prashant PM. Predicting aquatic ecosystem health using machine learning algorithms. International Journal of Aquatic Research and Environmental Studies. 2024;4(S1):39-44. https://doi.org/10.70102/IJARES/V4S1/7

[6] Mehrpour O, Hoyte C, Delva-Clark H, Al Masud A, Biswas A, Schimmel J, Nakhaee S, Goss F. Classification of acute poisoning exposures with machine learning models derived from the National Poison Data System. Basic & clinical pharmacology & toxicology. 2022 Dec;131(6):566-74. https://doi.org/10.1111/bcpt.13800

[7] Balamanikandan A, Saravanakumar M, Gunasekaran S, Anjum V, Gurusamy P, Ashokkumar N. Deep learning in the detection of chronic kidney disease. In2023 4th International Conference on Intelligent Technologies (CONIT) 2024 Jun 21 (pp. 1-6). IEEE. https://doi.org/10.1109/CONIT61985.2024.10627434

[8] Rao BK, Kumar PS, Reddy DK, Nayak J, Naik B. QCM sensor-based alcohol classification by advance machine learning approach. In Intelligent Computing in Control and Communication: Proceeding of the First International Conference on Intelligent Computing in Control and Communication (ICCC 2020) 2021 Jan 5 (pp. 305-320). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-8439-8_25

[9] Dharmireddi S, Mahdi HM, Rajendran M, Suryasa IW, Soy A. Artificial Intelligence-Driven Natural language processing for the futuristic Language Processing. In2025 International Conference on Computational Innovations and Engineering Sustainability (ICCIES) 2025 Apr 24 (pp. 1-6). IEEE. https://doi.org/10.1109/ICCIES63851.2025.11033144

[10] Ebrahimi A, Wiil UK, Andersen K, Mansourvar M, Nielsen AS. A predictive machine learning model to determine alcohol use disorder. In2020 IEEE Symposium on Computers and Communications (ISCC) 2020 Jul 7 (pp. 1-7). IEEE. https://doi.org/10.1109/ISCC50000.2020.9219685

[11] Patil BM, Joshi RC, Toshniwal D. Association rule for classification of type-2 diabetic patients. In2010 second international conference on machine learning and computing 2010 Feb 9 (pp. 330-334). IEEE. https://doi.org/10.1109/ICMLC.2010.67

[12] Ordonez C. Comparing association rules and decision trees for disease prediction. InProceedings of the international workshop on Healthcare information and knowledge management 2006 Nov 11 (pp. 17-24). https://doi.org/10.1145/1183568.1183573

[13] Rodrigues D, Ribeiro G, Siqueira V, Costa RM, Barbosa R. Associative patterns in health data: exploring new techniques. Health and Technology. 2022 Mar;12(2):415-31. https://doi.org/10.1007/s12553-021-00635-6

[14] Rashid MA, Hoque MT, Sattar A. Association rules mining based clinical observations. arXiv preprint arXiv:1401.2571. 2014 Jan 11. https://doi.org/10.48550/arXiv.1401.2571

[15] Narins RG, Emmett M. Simple and mixed acid-base disorders: a practical approach. Medicine. 1980 May 1;59(3):161-82.

[16] Kraut JA, Madias NE. Serum anion gap: its uses and limitations in clinical medicine. Clinical journal of the American Society of Nephrology. 2007 Jan 1;2(1):162-74. https://doi.org/10.2215/CJN.03020906

[17] Ponnarengan H, Rajendran S, Khalkar V, Devarajan G, Kamaraj L. Data-Driven Healthcare: The Role of Computational Methods in Medical Innovation. CMES-Computer Modeling in Engineering & Sciences. 2025 Jan 1;142(1). https://doi.org/10.32604/cmes.2024.056605

[18] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019 Apr 4;380(14):1347-58. https://doi.org/10.1056/NEJMra1814259

[19] Ozdemir H, Sasmaz MI, Guven R, Avci A. Interpretation of acid–base metabolism on arterial blood gas samples via machine learning algorithms. Irish Journal of Medical Science (1971-). 2025 Feb;194(1):277-87. https://doi.org/10.1007/s11845-024-03767-6

[20] Gün M. AI-assisted blood gas interpretation: a comparative study with an emergency physician. The American Journal of Emergency Medicine. 2025 Apr 14. https://doi.org/10.1016/j.ajem.2025.04.028

[21] Alrashed M, Aldeghaither NS, Almutairi SY, Almutairi M, Alghamdi A, Alqahtani T, Almojathel GH, Alnassar NA, Alghadeer SM, Alshehri A, Alnuhait M. The perils of methanol exposure: insights into toxicity and clinical management. Toxics. 2024 Dec 20;12(12):924. https://doi.org/10.3390/toxics12120924

[22] Ahmadi S, Ostadi A, Chitsazi H, Alikhah H. Clinical and laboratory prognostic factors associated with methanol toxicity outcomes in patients at Tabriz Sina Hospital: A retrospective study. Human & Experimental Toxicology. 2025 Jul 1;44:09603271251358632. https://doi.org/10.1177/09603271251358632

[23] Guy C, Holmes NE, Kishore K, Marhoon N, Serpa-Neto A. Decompensated metabolic acidosis in the emergency department: Epidemiology, sodium bicarbonate therapy, and clinical outcomes. Critical Care and Resuscitation. 2023 Jun 1;25(2):71-7. https://doi.org/10.1016/j.ccrj.2023.05.003

[24] AlSamh DA, Kramer AH. Neurologic complications in critically Ill patients with toxic alcohol poisoning: A multicenter population-based cohort study. Neurocritical Care. 2024 Apr;40(2):734-42. https://doi.org/10.1007/s12028-023-01821-2

[25] Ge Y, Ma Y, Lv P, Ren J, Wang Z, Zhang C. Association between albumin-corrected anion gap and delirium in acute pancreatitis: insights from the MIMIC-IV database. BMC gastroenterology. 2025 Aug 5;25(1):554. https://doi.org/10.1186/s12876-025-04150-0

[26] Caballé-Cervigón N, Castillo-Sequera JL, Gómez-Pulido JA, Gómez-Pulido JM, Polo-Luque ML. Machine learning applied to diagnosis of human diseases: A systematic review. Applied Sciences. 2020 Jul 26;10(15):5135. https://doi.org/10.3390/app10155135

[27] Mousavinejad SN, Lachouri R, Bahadorzadeh M, Khatami SH. Artificial intelligence for arterial blood gas interpretation. Clinica Chimica Acta. 2025 Oct 29:120691. https://doi.org/10.1016/j.cca.2025.120691

[28] Fan T, Wang J, Li L, Kang J, Wang W, Zhang C. Predicting the risk factors of diabetic ketoacidosis-associated acute kidney injury: A machine learning approach using XGBoost. Frontiers in public health. 2023 Apr 6;11:1087297. https://doi.org/10.3389/fpubh.2023.1087297

[29] Hassan MR, Huda S, Hassan MM, Abawajy J, Alsanad A, Fortino G. Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion. Information Fusion. 2022 Jan 1;77:70-80. https://doi.org/10.1016/j.inffus.2021.07.010

[30] Chen ML, Jiao Y, Fan YH, Liu YH. Artificial intelligence for early prediction of alcohol-related liver disease: Advances, challenges, and clinical applications. Artificial Intelligence in Gastroenterology. 2025 Jun 8;6(1). http://dx.doi.org/10.35712/aig.v6.i1.107193

[31] Martono NP, Kuramaru S, Igarashi Y, Yokobori S, Ohwada H. Blood alcohol concentration screening at emergency room: Designing a classification model using machine learning. In2023 14th International Conference on Information & Communication Technology and System (ICTS) 2023 Oct 4 (pp. 255-260). IEEE. https://doi.org/10.1109/ICTS58770.2023.10330879

[32] Ghazi A, Alisawi M, Hammood L, Abdullah SS, Al-Dawoodi A, Ali AH, Almallah AN, Hazzaa NM, Wahab YM, Nawaf AY. Data mining and machine learning techniques for coronavirus (COVID-19) pandemic: A review study. InAIP Conference Proceedings 2023 Sep 29 (Vol. 2839, No. 1, p. 040010). AIP Publishing LLC.

[33] Mehrpour O, Hoyte C, Delva-Clark H, Al Masud A, Biswas A, Schimmel J, Nakhaee S, Goss F. Classification of acute poisoning exposures with machine learning models derived from the National Poison Data System. Basic & clinical pharmacology & toxicology. 2022 Dec;131(6):566-74. https://doi.org/10.1111/bcpt.13800

[34] Ghazi A, Alisawi M, Hammood L, Abdullah SS, Al-Dawoodi A, Ali AH, Almallah AN, Hazzaa NM, Wahab YM, Nawaf AY. Data mining and machine learning techniques for coronavirus (COVID-19) pandemic: A review study. InAIP Conference Proceedings 2023 Sep 29 (Vol. 2839, No. 1, p. 040010). AIP Publishing LLC. https://doi.org/10.1063/5.0167882

[35] Olson DL, Araz ÖM. Data mining and analytics in healthcare management. International Series in Operations Research &Management Science. 2023.

[36] Mahmood AH, Al-Awadi SJ, Al-Attar MM, Alshammary RA, Abood RS. Investigate the association between genetic polymorphisms of ACE and ACE-2 with some biomarkers in Iraqi patients with COVID-19. Human Gene. 2024 Dec 1;42:201344. https://doi.org/10.1016/j.humgen.2024.201344

[37] Sánchez-de-Madariaga R, Martinez-Romo J, Escribano JM, Araujo L. Semi-supervised incremental learning with few examples for discovering medical association rules. BMC medical informatics and decision making. 2022 Jan 24;22(1):20. https://doi.org/10.1186/s12911-022-01755-3

[38] Periyasamy S, Kaliyaperumal P, Thirumalaisamy M, Balusamy B, Elumalai T, Meena V, Jadoun VK. Blockchain enabled collective and combined deep learning framework for COVID19 diagnosis. Scientific Reports. 2025 May 13;15(1):16527. https://doi.org/10.1038/s41598-025-00252-7