



Drug Addiction Likelihood Prediction Using a Web-Based Platform: (RiskAssess)

Mohamed Shabri ^a, N.M. Hakmanage ^b

^a*Department of Information Technology, Faculty of Computing CINEC Campus (Pvt) Ltd.*

^b*Faculty of Computing and Technology, University of Kelaniya, Sri Lanka.*

^a*Email: shabrimohamed00@gmail.com*

^b*Email: navodim@kln.ac.lk*

Abstract

This paper presents RiskAssess as an early feasibility study, a comprehensive web-based platform that integrates machine learning algorithms with clinical workflow to predict drug addiction risk in the Sri Lankan context. The platform addresses the critical gap between risk identification and intervention by providing immediate counselor assignment, appointment scheduling, and secure messaging capabilities. A multi-step questionnaire covering demographics, environment, relationships, substance use patterns, and mental health was developed and deployed in English, Sinhala, and Tamil. Eight machine learning algorithms were evaluated, with XGBoost, KNN, and ANN achieving the highest accuracy of 90.91%. A hybrid ensemble model combining Random Forest, XGBoost, and SVM was implemented for production, achieving 81.82% accuracy with superior stability and calibrated probability outputs. Feature importance analysis identified motivation to stay sober (17.5%), alcohol use (8.7%), and drug cravings (5.7%) as the most significant predictors. The platform features role-based interfaces for clients, counselors, and administrators, enabling seamless integration from assessment to intervention. Comprehensive security measures including bcrypt password hashing, role-based access control, and audit logging ensure compliance with healthcare data protection standards. The system successfully demonstrates how locally adapted machine learning models can enhance early intervention strategies for substance use disorders while maintaining clinical workflow integration. However, the results are preliminary due to the small dataset size (n=102) and the platform requires further validation before clinical deployment.

Keywords: Machine learning; drug addiction prediction; risk assessment; web platform; sri lanka; ensemble

1. Introduction

Substance misuse is a serious ongoing problem in the population health worldwide and an increase in its prevalence is being witnessed in the developing nation, Sri Lanka, mostly in the adolescent age group. The traditional evaluation methods rely on self-assessment questionnaires (paper-based) and clinical interview guidelines; these processes could significantly expose potential sufferers of the disorder to the treatment gap. The situation is also complicated by the fact that the stigma of the substance-related disorders remains in place and thus people with the disorders tend to approach with help only when it is too late. (Shahriar et al., 2019).

With the emergence of the modern machine learning technologies, the fine-grained prediction of the risk of addictions becomes the possibility under the data-driven strategies. Most of the available current systems, however, are designed to fit the Western environment and thus are not tied into the localised healthcare logistics. This deficiency is addressed in the present study through development of RiskAssess as a streamlined, web-based Risk Assessment application that is specifically tailored to the environment in Sri Lanka.

Current literature lacks integrated platforms that combine locally adapted risk assessment with real-time clinical workflow integration for South Asian populations. This study aims to: (1) develop a culturally tailored assessment tool for Sri Lankan demographics, (2) evaluate machine learning approaches for addiction risk prediction, (3) integrate seamless counselor assignment workflows, and (4) demonstrate proof-of-concept feasibility while acknowledging limitations for future validation.

RiskAssess is a digital package that combines a locally tuned survey with machine-learning and enables rapid evidence-based risk categorization and providing an immediate clinical linkage to counseling services. The unique contribution of this work lies in the Sri Lankan cultural adaptation of questionnaires, integration of Sinhala and Tamil language support, and seamless workflow connecting assessment to counselor assignment - features not addressed in existing literature. The idea behind the system is to address one continuous issue in healthcare practice, and that is time between risk identification and response. The platform will also allow healthcare professionals to provide preventive assistance to people at a higher risk of developing substance use disorders because the platform will enable healthcare providers to define the risks in real-time. The innovation in it lies in the fact that it eliminates the administrative risks of traditional health-care environment and smoothenes the process between initial assessment and treatment, as well as reduces the time delay between risk identification and its actual mitigation. This work presents an early proof-of-concept requiring further validation with larger datasets. (Kaur & Bawa, 2017).

2. Methodology

2.1 System Architecture

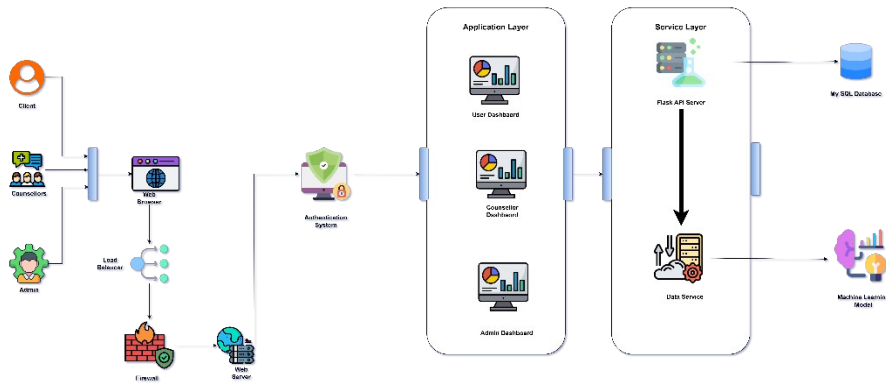


Figure 1: System Architecture Diagram

RiskAssess implements three-tier architecture where presentation and application layers are selected and defined. The presentation layer uses HTML, CSS, JavaScript, and PHP to create an interface that is specific to the role of the user that is of the clients, the counselor, and administrators of the system (Shen Haibo & Hong Fan, 2005). The application has a business logic wrapped in it and the Flask API is used to interact with an embedded machine-learning model. The data layer is in charge of persistent user data, examinations, appointments, and audit logs, with the exception of customer documents, and it uses MySQL

2.2 Data Collection and Preprocessing

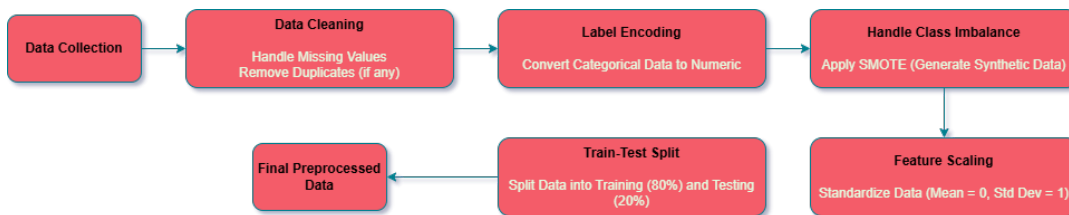


Figure 2: Steps of data preprocessing of gathered data

Data was gathered on a questionnaire carried out on Google forms in English, Sinhala and Tamil. The questionnaire included 43 variables included into five domains namely, demographics, environment, relationships, pattern of substance use, and mental health indicators, and was self-administered by the residents of Sri Lanka, with and without a history of illicit drug use (Mityagin et al., 2014).

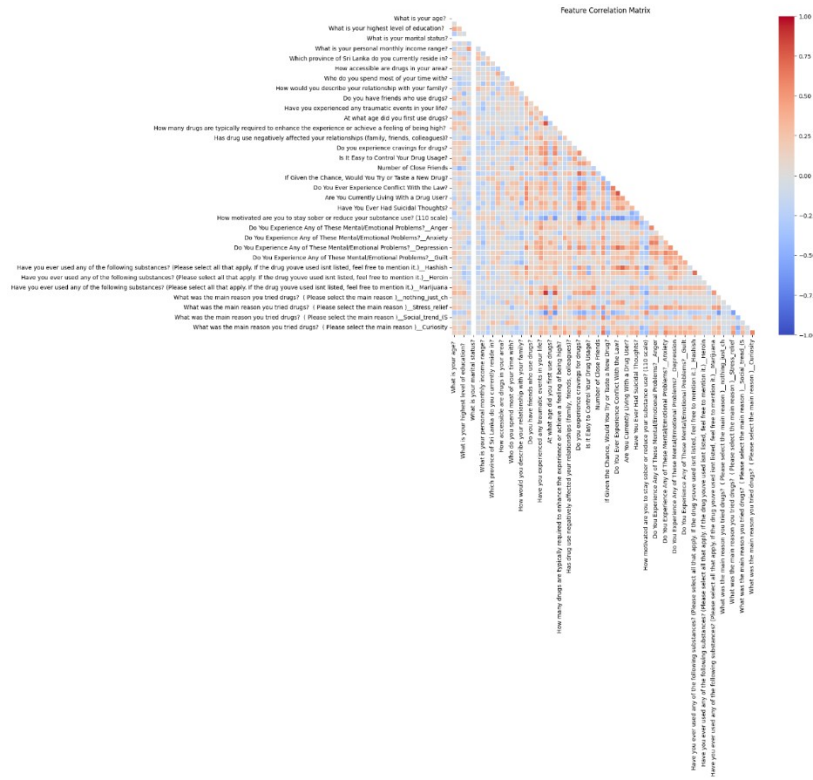


Figure 3: Correlation Matrix

The preprocessing pipeline included handling missing values through mean imputation for numerical features and mode imputation for categorical variables (Harsha et al., 2022). Multiple-choice responses were transformed into binary indicator variables using one-hot encoding. Feature selection was performed using ANOVA F-tests, followed by standardization and Principal Component Analysis (PCA) retaining 95% of variance (Deepa, 2023). Class imbalance was addressed using Synthetic Minority Oversampling Technique (SMOTE) (Bharathidasan & Sujdha, 2023).

2.3 Machine Learning Model Development

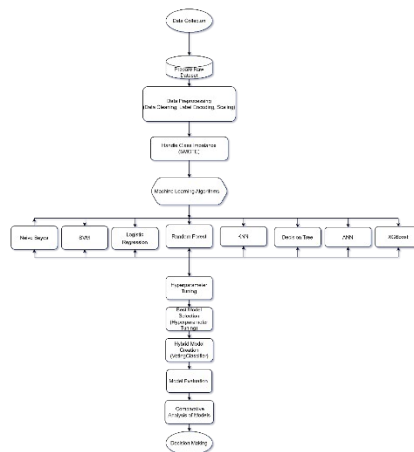


Figure 4: Machine Learning Model Development

Eight algorithms were evaluated: Naive Bayes, Support Vector Machine (SVM), Logistic Regression, XGBoost, Random Forest, K-Nearest Neighbors (KNN), Decision Tree, and Artificial Neural Network (ANN) (Islam Arif et al., 2021). Model performance was assessed using stratified 5-fold cross-validation with accuracy, precision, recall, and F1-score metrics.

A risk stage creation methodology was implemented using a weighted scoring system based on clinical literature. Drug usage frequency contributed 0-80 points, motivation to stay sober (inverse relationship) contributed 0-50 points, while cravings, withdrawal symptoms, and control over usage each contributed 20 points. Scores below 40 indicated Low risk, 40-70 Moderate risk, and above 70 High risk.

Hyperparameter optimization was conducted using GridSearchCV for Random Forest and XGBoost models. Probability calibration was applied to SVM (Abada & Bouramoul, 2022), Random Forest, and XGBoost using CalibratedClassifierCV to ensure reliable probability estimates for clinical decision-making (Rashedul Alam et al., 2021).

A hybrid ensemble model combining Random Forest, XGBoost, and SVM was developed using soft voting to leverage the strengths of multiple algorithms while improving generalization and stability.

3. Results

3.1 Model Performance Analysis

Figure captions and table headings should be sufficient to explain the figure or table without needing to refer to the text. Figures and tables not cited in the text should not be presented. The following is an example for Table 1.

Table 1. Machine Learning Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score
Naive Bayes	72.73	0.79	0.67	0.69
SVM	63.64	0.71	0.58	0.59
Logistic Regression	81.82	0.88	0.75	0.78
XGBoost	90.91	0.93	0.83	0.85
Random Forest	72.73	0.67	0.67	0.66
KNN	90.91	0.94	0.92	0.92
Decision Tree	54.55	0.57	0.52	0.50
ANN	90.91	0.94	0.92	0.92
Hybrid Model	81.82	0.86	0.75	0.78

XGBoost, KNN, and ANN achieved the highest accuracy of 90.91%, with KNN and ANN demonstrating superior precision and recall scores. However, the hybrid ensemble model was selected for implementation due to its superior stability across different data partitions and calibrated probability outputs essential for clinical applications.

While the reported accuracy appear promising, the small dataset size (n=102) presents significant limitations. The high accuracy scores may be overestimated due to potential overfitting, and the lack of external validation limits generalizability. Future studies require larger, more diverse datasets and independent test sets to validate these preliminary findings.

3.2 Cross-Validation Results

Stratified 5-fold cross-validation revealed that while individual models achieved higher test accuracy, Logistic Regression and Random Forest showed the most consistent performance with mean accuracy of 0.850. XGBoost demonstrated the lowest standard deviation (0.033), indicating high stability regardless of data partitioning.

3.3 Feature Importance Analysis

Feature importance analysis using Random Forest identified the most significant predictors of addiction risk:



Figure 5. Top Predictive Features

3.4 System Implementation Results

The platform brings risk prediction and clinical work together in a smooth and practical way. High-risk cases are identified in real time, and counselors are notified immediately. Appointments are set up automatically, depending on when counselors are free. A secure messaging system makes it easy for clients and counselors to stay in touch.

All actions are recorded in detail to meet safety and compliance standards. Access to information is managed based on each user’s role, so only the right people see the right data.

RiskAssess

Dashboard

Profile

Book Appointment

My Appointments

Messages

Predict Drug Addiction Likelihood

Welcome, User!

Your RiskAssess dashboard

Manage your counselling appointments and personal information here.

Substance Use Risk Assessment

Confidential multi-step screening to understand your risk factors

Demographics & Background

Basic information about yourself

What is your age?

Select...

What is your gender?

Select...

What is your highest level of education?

Select...

What is your current employment status?

Select...

What is your marital status?

Select...

What is your family's income level?

Select...

What is your personal monthly income range?

Select...

Figure 6. Risk Assessment Form Client Dashboard

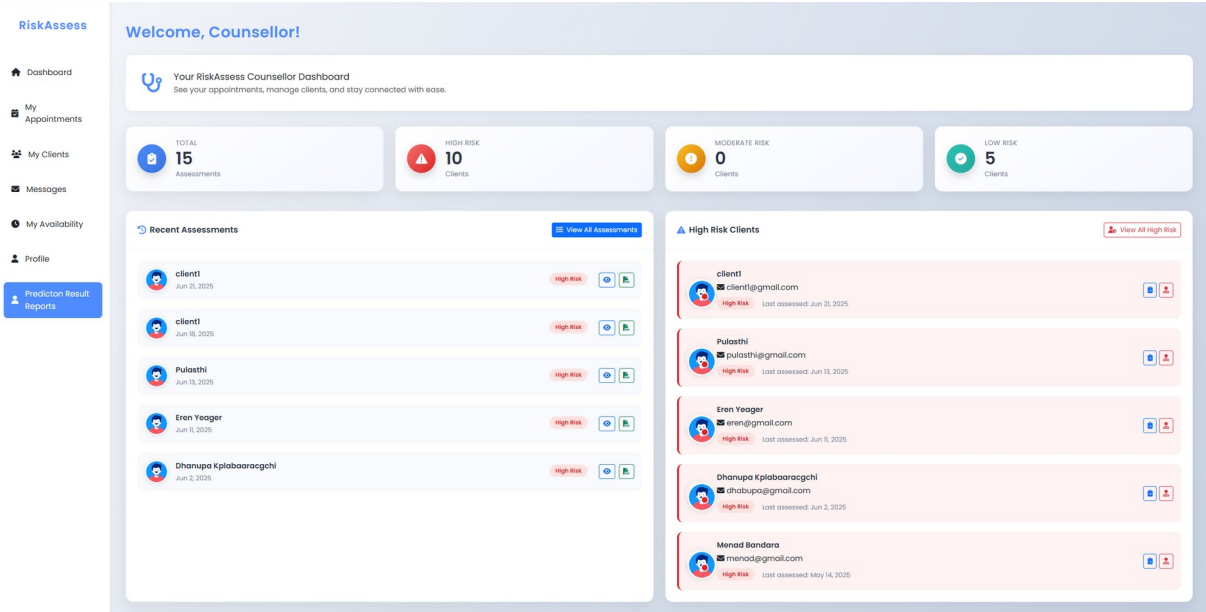


Figure 7. Counsellor Dashboard



Figure 8. Counsellor Dashboard Risk Distribution Analysis

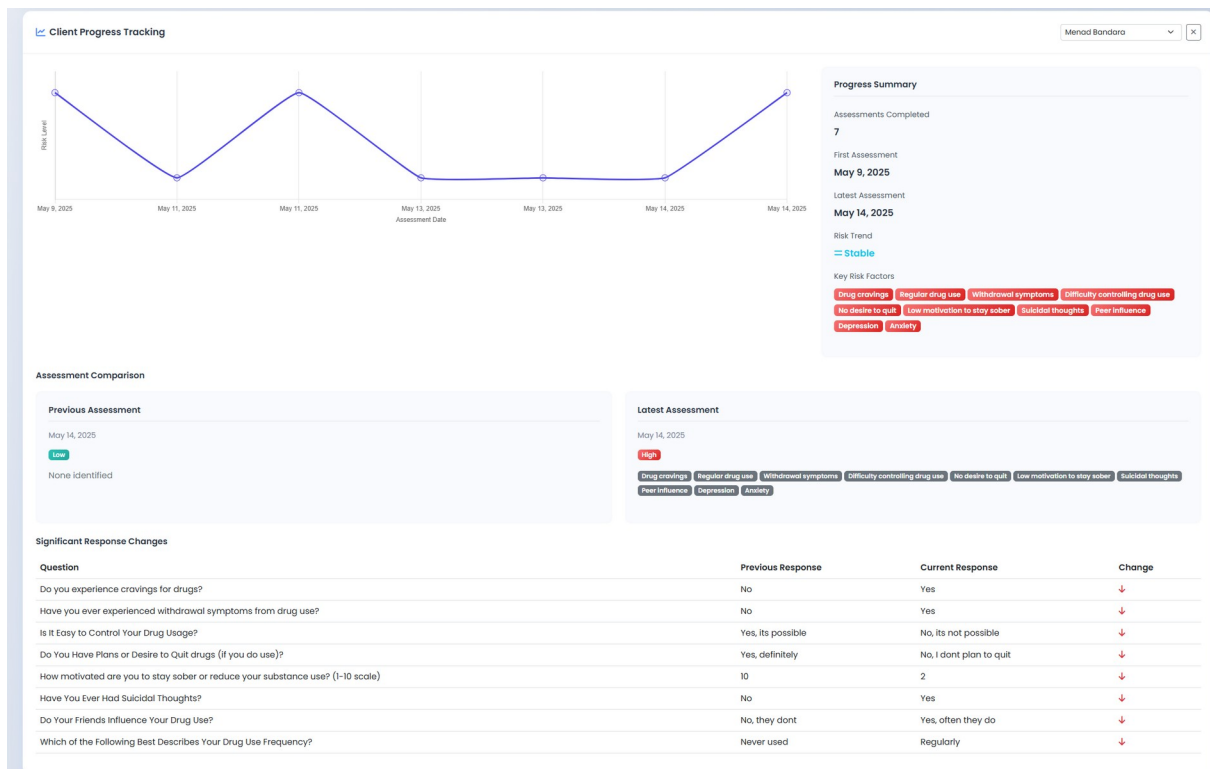


Figure 9. Counsellor Dashboard Client Progress Tracking

4. Discussion

The results help to demonstrate that the application of machine learning in combination with the locally customized questionnaires and clinical workflows can accurately identify addiction risk. Even though the hybrid ensemble system does not stand the best single system accuracy, it has the stability and reliability which is necessary in the field of healthcare usage.

The fact that motivation to remain sober was the key predictor supports the claim of the significant role of psychological variables to determine the risk of addiction. These findings guide the treatment that focuses on motivational change and behavioral change and thus strengthens the results of treatment.

The platform's integration of assessment, prediction, and intervention represents a significant advancement over traditional paper-based methods. By providing immediate risk classification and counselor assignment, the system addresses the critical gap between identification and treatment initiation.

The broad structure of security systems, which include encrypted data transmissions, role based access and audit logs, can be used to meet the current requirements of data-protection standards within the healthcare profession and does not compromise the usability of the system.

4.1 Limitations and Future Work

The relatively small dataset size (n=102) significantly affects model reliability and generalizability. The risk of overfitting is substantial, and model performance metrics may be overestimated. The absence of external validation data prevents assessment of real-world performance.

The study lacks independent test data from different populations or time periods. Cross-validation alone is insufficient for clinical validation. The self-reported nature of data introduces potential bias.

This work represents an early feasibility study and should not be used for clinical decision-making without extensive validation. Real-world testing in clinical environments is essential before deployment.

The platform currently supports only English in counselor and administrative interfaces. Multi-language support would enhance accessibility across Sri Lanka's diverse linguistic landscape. Integration with existing healthcare information systems and the development of mobile applications would further improve platform accessibility and adoption.

Immediate priorities include expanding the dataset to include more diverse populations (target n>1000), conducting external validation studies, performing clinical trials with healthcare professionals, and developing comprehensive safety protocols for clinical deployment.

5. Conclusion

This early feasibility study demonstrates the potential for combining machine learning with clinical workflow in Sri Lankan addiction risk assessment. The locally adapted platform design, integrated with ensemble machine learning approaches, shows promise for risk estimation while ensuring seamless intervention service integration.

The system's direct counselor assignment and integrated appointment scheduling addresses critical gaps in traditional assessment procedures. The culturally adapted questionnaires and multi-language support represent novel contributions to South Asian healthcare technology.

Results are preliminary and limited by small sample size (n=102), potential overfitting, and lack of external validation. The platform is not clinically validated and requires extensive testing before healthcare deployment.

This research provides foundational evidence for machine learning applications in healthcare, particularly for early intervention in substance use disorders. The modular architecture offers a framework for expansion to other mental health conditions, contingent on proper validation studies.

These results should not be used for clinical decision-making without further validation through larger studies and clinical trials.

Acknowledgements

The support of CINEC Campus Faculty of Computing is acknowledged for providing the necessary resources and infrastructure for this research. Special thanks are extended to all participants who contributed data for model development and to the project supervisor for valuable guidance and mentorship during the research process.

References

- Abada, W., & Bouramoul, A. (2022). Using machine learning techniques to predict people at-risk for drug addiction: A Bayesian-based model. In *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)* (pp. 1-7). IEEE. <https://doi.org/10.1109/PAIS56586.2022.9946914>
- Bharathidason, D. S., & Sujdha, C. (2023). A comparative study of machine learning algorithms for drug addiction prediction. *International Journal of Advanced Computer Science and Applications*, 44(12).
- Deepa, D. M. (2023). Drug addiction prediction system by machine learning techniques: A case study. *International Journal of Computer Applications*, 8(2).
- Harsha, K., Banarjee, K., Vinooth, P., Singh, S., Tiwari, V., & Shukla, V. (2022). Prediction of narcotics addiction: A review. In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 2211-2216). IEEE. <https://doi.org/10.1109/IC3I56241.2022.10073168>
- Islam Arif, M. A., Sany, S. I., Sharmin, F., Rahman, M. S., & Habib, M. T. (2021). Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population. *International Journal of Electrical and Computer Engineering*, 11(5), 4471-4480. <https://doi.org/10.11591/ijece.v11i5.pp4471-4480>
- Kaur, S., & Bawa, R. K. (2017). Implementation of an expert system for the identification of drug addiction using decision tree ID3 algorithm. In *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-6). IEEE. <https://doi.org/10.1109/ICACCAF.2017.8344677>
- Kumara, U. G. H. T., Siriwardana, S. S. A., Weerasinghe, L., Shavindi, R. A. K. I., Chiranjeeva, H. P. R. C., & Siriwardana, S. (2023). A machine learning approach to analyze drug addiction. In *2023 5th International Conference on Advancements in Computing (ICAC)* (pp. 113-118). IEEE. <https://doi.org/10.1109/ICAC60630.2023.10417256>
- Mityagin, S. A., Ivanov, S. V., & Boukhanovsky, A. V. (2014). Multi-factorial predictive modelling of drug addiction for large urban areas. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-7). IEEE. <https://doi.org/10.1109/ICAICT.2014.7035978>
- Rashedul Alam, K. M., Ahammed, K., Rony, M. A. T., & Ferdousi, Z. (2021). A comparative machine learning study to predict drug addiction in Bangladesh. In *2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-6). IEEE. <https://doi.org/10.1109/AICT52784.2021.9620453>

Shahriar, A., Faisal, F., Mahmud, S. U., Chakrabarti, A., & Rabiul Alam, M. G. (2019). A machine learning approach to predict vulnerability to drug addiction. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-7). IEEE.

<https://doi.org/10.1109/ICCIT48885.2019.9038605>

Shen, H., & Hong, F. (2005). A context-aware role-based access control model for web services. In *IEEE International Conference on E-Business Engineering (ICEBE)* (pp. 220-223). IEEE.

<https://doi.org/10.1109/ICEBE.2005.1>