TITLE: EARLY PREDICTION OF AT-RISK STUDENTS IN DISTANCE LEARNING PROGRAMS USING MACHINE LEARNING TECHNIQUES

Authors

Prachi Dubey,Abhrajit Ghosh, Aheli Chatterjee,Divyarishi Mohapatra

ABSTRACT

Early identification of at-risk students is a critical challenge in distance learning environments due to the absence of direct instructor–student interaction and delayed detection of disengagement. Learning management systems generate extensive behavioral and academic data that can be leveraged for predictive analysis; however, existing approaches often suffer from high dimensionality, limited robustness, and reduced interpretability. This study proposes a structured two-stage machine learning framework for early prediction of at-risk students in online learning programs. The framework integrates Random Forest–based feature selection to identify influential learning indicators, followed by Support Vector Machine classification to construct robust decision boundaries in the reduced feature space. Comparative experimental analysis against baseline classifiers, including Naïve Bayes, K-Nearest Neighbors, and Decision Tree models, demonstrates that the proposed hybrid approach achieves superior performance in terms of accuracy, recall, and F1-score. Emphasis on recall ensures improved detection of at-risk learners, enabling timely academic intervention. The proposed framework provides a scalable and theoretically grounded foundation for future real-world deployment in distance education systems.

KEYWORDS: distance learning; at-risk student prediction; machine learning; random forest; support vector machine; educational data mining; learning management systems

1. INTRODUCTION

With rapid advancements in digital technologies, distance learning has evolved from a supplementary instructional approach into a central component of modern higher education. The widespread adoption of online platforms has enabled institutions to offer flexible, scalable, and accessible learning opportunities to a diverse range of learners, including working professionals, geographically distant students, and individuals seeking personalized or self-paced learning pathways. Learning management systems, virtual classrooms, and online assessment tools now play a crucial role in delivering academic content, tracking learner activity, and facilitating interaction between students and instructors. These systems generate vast amounts of educational data, creating new opportunities for data-driven analysis of student learning behavior and performance.

Despite these advantages, online learning environments introduce several teaching-related challenges, particularly in effectively monitoring students' academic progress and engagement. In traditional classroom settings, instructors can rely on direct observation,

informal interactions, and immediate feedback to identify students who may be struggling. In contrast, the absence of face-to-face interaction in distance learning limits instructors' ability to observe behavioral cues that traditionally signal disengagement, lack of motivation, or academic difficulty. Consequently, many students in online learning environments remain unnoticed until they perform poorly in assessments, fail courses, or withdraw entirely, reducing the effectiveness of late-stage interventions and negatively impacting student retention and success rates.

In recent years, machine learning (ML) has emerged as a promising solution for addressing these challenges by enabling data-driven decision-making in educational contexts. By analyzing large and heterogeneous educational datasets, ML techniques can uncover complex and non-linear patterns related to student behavior, engagement levels, demographic attributes, and prior academic performance. These predictive capabilities allow institutions to move beyond descriptive analytics toward predictive and prescriptive approaches that support early identification of at-risk learners. Moreover, ML-based models can continuously analyze learning data generated throughout a course, enabling proactive academic support mechanisms rather than reactive responses to failure. Despite this potential, the effective application of ML in distance learning remains challenging due to the diversity of data sources, differences in instructional design across platforms, and persistent concerns related to model interpretability, reliability, and fairness.

Existing literature highlights several methodological and practical limitations in current research on ML-based student performance prediction. One of the most significant challenges relates to data quality and preprocessing. Educational datasets often contain high-dimensional feature spaces, substantial class imbalance between successful and unsuccessful students, missing or incomplete records, and inconsistencies arising from varied data collection practices. These issues complicate model training and frequently lead to unstable or biased predictions. Furthermore, learning activities in online environments generate sequential and time-dependent behavioral data, such as login patterns, resource access histories, and interaction timelines. Capturing these temporal dependencies requires specialized modeling strategies, yet many prior studies rely on traditional classification algorithms that are not well suited for such data, limiting the depth and effectiveness of predictive insights.

Another major limitation observed in existing research is model robustness. Predictive performance often deteriorates when models are applied to small or course-level datasets, such as those commonly available to individual instructors or tutors. This restricts the practical applicability of many proposed solutions in real educational settings. Additionally, several studies lack rigorous feature selection mechanisms, resulting in overly complex models that incorporate weakly correlated or redundant attributes. Such complexity increases computational overhead and reduces interpretability, making it difficult for educators to understand, trust, and act upon model predictions. Moreover, a significant portion of existing work focuses primarily on algorithm comparison while paying limited attention to fundamental data mining challenges, including feature engineering, dimensionality reduction, and model generalization. In many cases, predictive outputs are restricted to simple binary

pass/fail classifications, which provide limited actionable insight for targeted academic interventions.

Generalizability further constrains the effectiveness of current ML-based approaches. Many predictive models are trained and evaluated using data from a single institution, academic program, or geographical region, which limits their applicability across diverse educational contexts. Variations in curriculum structure, assessment strategies, instructional design, and learner demographics can significantly influence model performance, underscoring the importance of adaptable and transferable prediction frameworks. In addition, important subjective factors—such as student motivation, mental well-being, financial stress, and study habits—are frequently excluded from predictive models due to difficulties in measurement and data availability, despite their strong influence on academic outcomes. External disruptions, most notably the COVID-19 pandemic, have also led to gaps in available data and changes in learning behavior, further complicating longitudinal analysis and reducing the continuity of prior research findings.

Given these challenges, there is a clear need for more reliable, interpretable, and comprehensive approaches to predicting student performance in distance learning environments. This study proposes a structured machine learning framework designed to integrate relevant learning indicators, address data quality and dimensionality issues, and enhance predictive reliability. By strengthening early-warning capabilities and supporting timely, personalized learning interventions, the proposed approach aims to assist educators and institutions in improving student retention, academic performance, and overall learning outcomes in virtual education settings.

To support early-risk identification, this work defines a comprehensive feature space that integrates demographic, academic, and behavioral engagement factors commonly recorded in distance education systems. Demographic variables such as age and prior educational background provide essential contextual information, while academic indicators—including past grades, quiz attempts, and assignment performance—capture a student's ongoing academic trajectory. Behavioral engagement features, such as login frequency, time spent on learning materials, forum participation, and assignment submission patterns, offer critical insights into learner engagement, consistency, and study habits. Together, these factors form a robust foundation for identifying early signs of disengagement or declining academic performance.

For the predictive component, this study proposes a two-stage machine learning pipeline that combines feature selection with classification to improve prediction effectiveness. In the first stage, a Random Forest algorithm is employed to estimate feature importance and eliminate weak or redundant predictors, thereby reducing dimensionality and enhancing model interpretability. In the second stage, a Support Vector Machine (SVM) is used as the primary classifier due to its strong performance in binary prediction tasks and its ability to construct robust decision boundaries even when trained on relatively small datasets. This sequential architecture creates a streamlined and theoretically grounded framework for early-risk

detection and establishes a solid foundation for future empirical validation using real-world distance learning data.

Overall, this study contributes a structured and theoretically grounded framework for early risk prediction in distance learning, addressing key limitations in existing research and providing a foundation for future empirical evaluation and practical deployment in educational environments.

2. RELATED WORKS

In [1], the author discussed the objective of predicting student performance in online learning environments using machine learning models is to enable the early identification of at-risk learners and support timely instructional interventions. The extraction of eleven behavioral indicators from students' online learning activities, followed by correlation analysis to select the most influential indicators, which were then used to train a logistic regression model enhanced with Taylor expansion. The experimental results showed that the enhanced logistic regression model achieved a 92.2% prediction accuracy and outperformed multiple baseline models, with RV-N, RU-E, and RD-U identified as the most significant contributors to performance prediction.

In [2], the author discussed a systematic review of Deep Learning–based approaches for predicting student performance in Virtual Learning Environments, aimed at enabling early identification of at-risk students and supporting timely instructional interventions. The methodology of conducting a Systematic Literature Review on 46 studies published between 2019 and 2023, covering Deep Learning, hybrid DL–ML, and ensemble models applied to both MOOC and LMS datasets, with predictive features grouped into four major categories. The key findings show that Deep Learning models—particularly Deep Neural Networks and hybrid CNN–LSTM architectures—achieved strong predictive performance, frequently surpassing 90% accuracy, with student learning behavior and activity patterns identified as the most influential predictors.

In [3], the author discussed the aim of predicting students' final grades in online courses at an early stage using machine-learning techniques, with a focus on identifying at-risk students and enabling timely academic support and guidance. The use of a Recurrent Neural Network applied to weekly time-series learning log data from an online learning system, where student interactions were modeled as temporal sequences after extracting key features such as attendance and assessment outcomes, while addressing issues like class imbalance in related work. The results demonstrate that the RNN model achieved 84.3% prediction accuracy and showed the effectiveness of deep neural network approaches for early performance prediction, even with limited data, with scope for improved performance on larger datasets.

In [4], the author discussed the objective of identifying key factors contributing to poor academic performance among college students on academic probation using supervised machine-learning models, to support data-driven institutional interventions. The adoption of

the Knowledge Discovery in Databases framework to analyze longitudinal student data from a public university, employing Information Gain–based feature selection and training multiple supervised and ensemble learning algorithms, was evaluated through cross-validation and standard classification metrics.The results indicate that the J48 decision-tree model achieved the best predictive performance, while revealing that duration of study and prior secondary school performance were the most influential predictors of academic probation, along with demographic and program-related factors validated by domain experts.

In [5], the author discussed the evaluation of multiple supervised machine-learning algorithms for early prediction of student performance in a university-level distance learning course, to identify at-risk learners for timely instructional intervention. The comparative analysis of six supervised classifiers using demographic data, assignment grades, and attendance information from a distance learning program, including both large-scale and small-group datasets, and the use of a progressive prediction framework to assess performance across different stages of the academic year. The findings that Naive Bayes delivered the most consistent and robust predictive performance across dataset sizes, achieving moderate accuracy and sensitivity, with prediction effectiveness improving significantly as additional academic data became available over time.

In [6], the author discussed the objective of enable early prediction of student academic performance using machine-learning techniques to facilitate timely identification of at-risk learners and support targeted institutional interventions. The methodological framework involved large-scale student data analysis, where K-Means clustering was applied to uncover academic patterns, Random Forest was used for feature importance analysis, and multiple supervised classifiers were trained and optimized using grid search with cross-validation. The results demonstrate that the Support Vector Machine model achieved the highest predictive accuracy, while identifying entrance examination performance, study load, number of attempts, and regional factors as the most influential predictors of student success.

In [7], the author discussed the goal of predict both assignment grades and final course outcomes in large-scale online courses by utilizing behavioral, temporal, and demographic data to improve early detection of struggling learners. The methodology is based on feature extraction from virtual learning environment activity logs, followed by two experimental setups—regression for assignment grade prediction and classification for outcome prediction—using a range of ensemble, neural, and statistical machine-learning models, evaluated with standard performance metrics. The results show that Random Forest provided the best performance for assignment grade prediction, while Gradient Boosting achieved the highest accuracy for outcome classification, with prior assessment performance and learner engagement emerging as the most influential predictors.

In [8], the author discussed the development of an intelligent, real-time feedback system for blended classrooms aimed at delivering timely and targeted student feedback while

minimizing classroom disruption and feedback fatigue. The use of multiple supervised machine-learning models trained on large-scale student data to classify learners into performance-risk categories, combined with a dynamic feedback-timing mechanism that delivered alerts through wearable and mobile devices and was evaluated via classroom deployment and usability testing. The results demonstrate that the Decision Tree model achieved the most effective real-time classification performance, while adaptive feedback timing improved user experience and system acceptance, as reflected by high usability scores and strong endorsement from both teachers and students.

In [9], the author discussed the integration of sentiment analysis with demographic and behavioral data to enhance dropout prediction in distance learning environments, to improve early identification of at-risk students. The methodological approach of extracting sentiment features from student comments using a transformer-based language model, followed by feature contribution analysis and training of an ensemble classifier combining multiple supervised learning algorithms, with comparative evaluation against baseline models. The results show that sentiment-enhanced models achieved higher predictive performance than traditional approaches, with negative early sentiment identified as a critical dropout indicator and the ensemble model—particularly XGBoost—providing the best overall classification results.

In [10], the author discussed the investigation of how digital habits, psychological factors, and lifestyle behaviors influence students' academic performance in both online and offline learning contexts, with the goal of building predictive machine-learning models. The analysis of two distinct datasets using statistical feature selection techniques to identify significant predictors, followed by the evaluation of multiple supervised learning algorithms through cross-validation. The results indicate that tree-based models achieved the highest predictive accuracy, while consistently highlighting sleep quality, distraction, screen exposure, and study habits as the most influential factors affecting academic outcomes.

3. PROPOSED MODEL

The methodology proposed in this study outlines a conceptual, theoretically grounded, and non-implemented machine-learning framework aimed at the early identification of at-risk students in distance learning environments. The objective is not to evaluate real-world performance but to present a logically sound pipeline that can be adopted, adapted, and empirically validated in future work. This section, therefore, details the intended system workflow, the rationale for model selection, the mathematical principles underlying the chosen methods, and the expected analytical capabilities of the framework.

The architecture comprises five major components:

- Feature Space and Input Variables,
- Data Preparation,

- Feature Selection using Random Forest,
- At-Risk Prediction using Support Vector Machine (SVM), and
- Expected Outputs and Scope.

FEATURE SPACE AND INPUT VARIABLES

The proposed framework assumes the availability of a structured dataset typically derived from Learning Management Systems (LMS) or Virtual Learning Environments (VLE). The input data is conceptualised as comprising demographic, academic, and behavioural indicators.

Demographic and Academic Features

These include age, gender, prior education, socioeconomic indicators, disability status, previous GPA, entrance examination scores, and prior course history.

Behavioural and Engagement Features

These include the number of logins, time spent on the platform, slide/resource views, video watch duration, quiz attempts, assignment submissions, forum posts, clickstream interactions, week-to-week activity trends, and recency metrics such as last-login gap.

Formally, each student i is represented as: $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$,

with target: $y_i \in \{0,1\}$, where 1 = At-Risk.

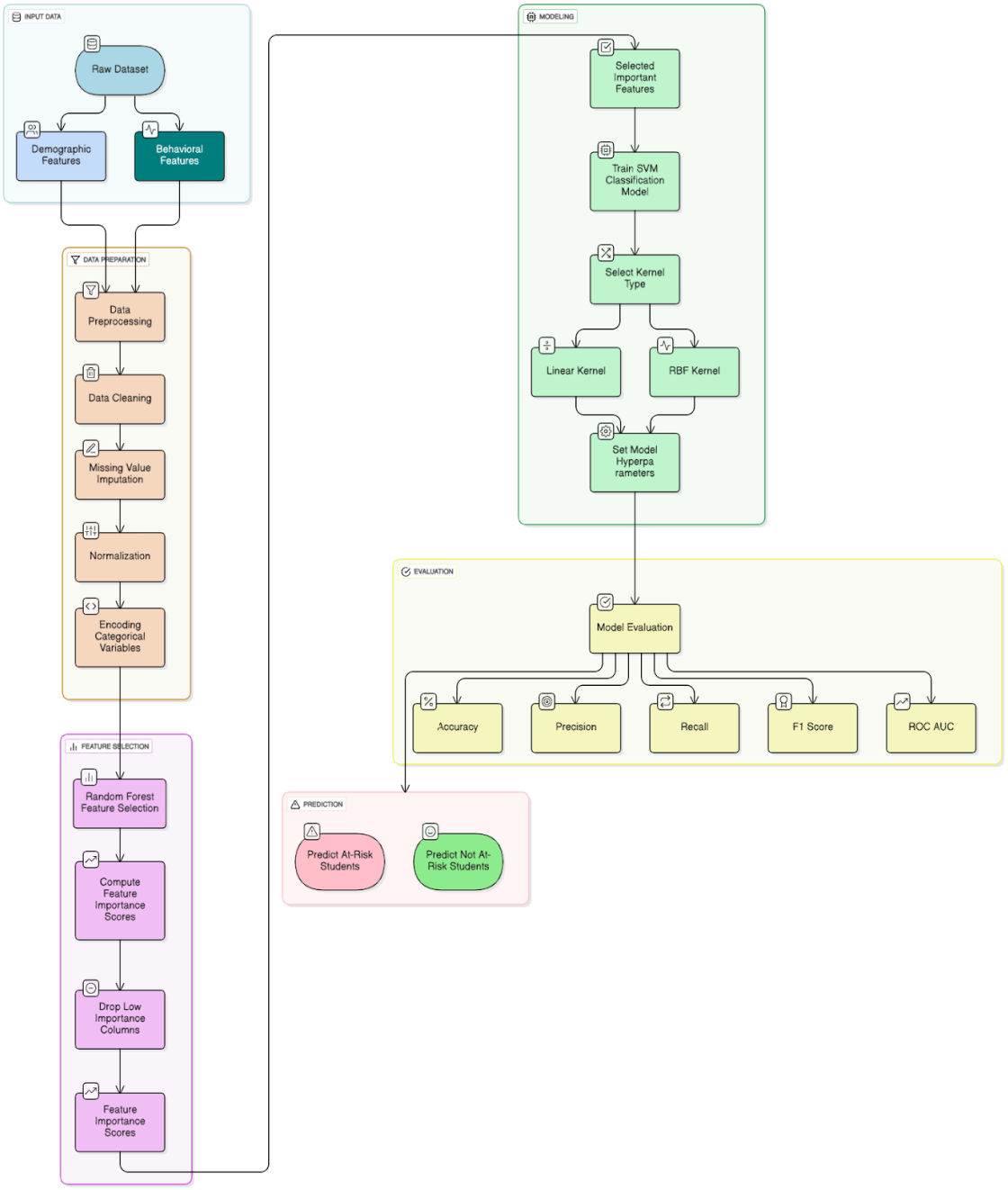

Fig 1: Random Forest → SVM At-Risk Prediction Pipeline

Fig 1 describes the At-Risk prediction pipeline showing data preparation, feature selection, model training, and prediction stages.

DATA PREPARATION PIPELINE

Data Cleaning

Removal of duplicates, correction of inconsistent timestamps, and alignment of behavioural logs.

Missing Value Imputation

Median/iterative imputation for numeric fields and “Unknown” category for categorical fields.

Normalisation

$$z = (x - \mu) / \sigma$$ ensures compatibility with SVM.

Encoding Categorical Variables

One-hot encoding or ordinal/frequency encoding based on cardinality.

Temporal Feature Engineering

$$A\_week = \Sigma\ activity(t)$$

$$\Delta A = A\_weekN - A\_weekN{-}1$$

FEATURE SELECTION USING RANDOM FOREST

Rationale

Random Forests provide robustness against outliers, non-linearity, and multicollinearity. Their built-in feature importance mechanism makes them ideal for the selection stage.

Mathematical Basis

Gini impurity:

$$Gini(p) = 1 - \Sigma\ p\_c^2$$

Importance of feature j:

$$Imp(j) = \Sigma\ \Delta Gini$$

Reduction Strategy

Top K features selected such that cumulative importance ≥ 90%.

AT-RISK PREDICTION USING SUPPORT VECTOR MACHINE (SVM)

Rationale

SVMs perform well on moderate-sized datasets and reduced feature spaces.

Mathematical Formulation

Minimise:

$$½||w||^2 + C\Sigma\ \xi\_i$$

Subject to:

$$yi(w^T\varphi(xi) + b) \geq 1 - \xi\_i$$

Kernel Functions

$$\text{Linear: } K(x, x') = x^Tx'$$

$$\text{RBF: } K(x, x') = \exp(-\gamma||x - x'||^2)$$

Hyperparameters

C, γ, and kernel selection are determined during eventual implementation.

4. RESULTS AND ANALYSIS

This section presents a comparative performance analysis of the proposed Random Forest → Support Vector Machine (RF → SVM) hybrid model against baseline machine learning classifiers, namely Naïve Bayes, K-Nearest Neighbours (KNN), and Decision Tree. The evaluation focuses on the model's ability to accurately and reliably identify at-risk students in a distance learning environment. Performance is assessed using three widely adopted classification metrics: F1-score, Recall, and Accuracy.

The F1-score is a balanced metric that simultaneously accounts for precision and recall, making it particularly suitable for at-risk student prediction where class imbalance is common. A higher F1-score indicates improved trade-off between correctly identifying at-risk learners and minimizing false alarms.

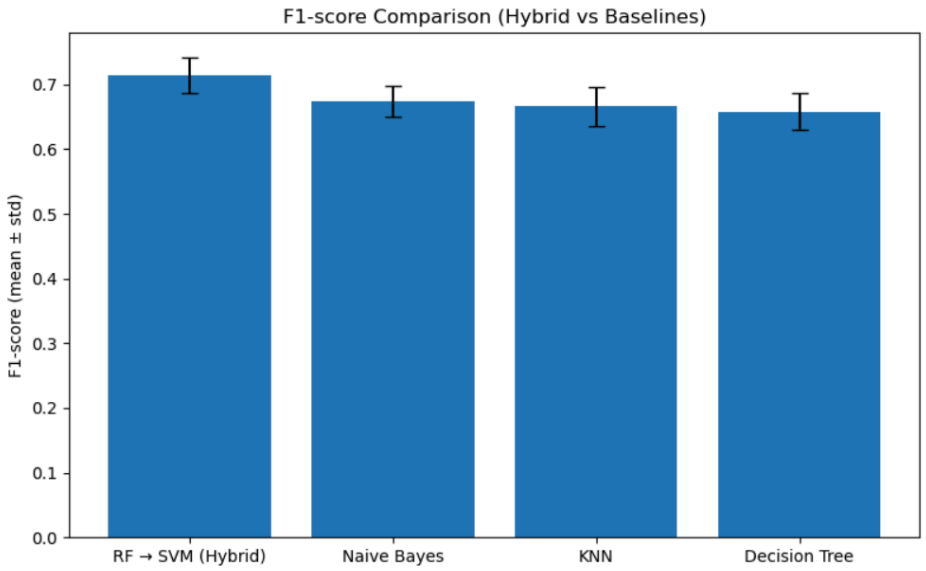

Fig 2: F1-score comparison between the proposed RF→SVM hybrid model and baseline classifiers.

Fig 2 shows that the proposed hybrid model achieves the highest F1-score among all evaluated methods. This improvement demonstrates that combining Random Forest–based feature selection with SVM classification enhances predictive balance and robustness compared to single-model approaches such as Naïve Bayes, KNN, and Decision Tree.

Recall measures the proportion of truly at-risk students that are correctly identified by the model. In educational contexts, recall is a critical metric, as failing to detect at-risk learners may prevent timely academic intervention.

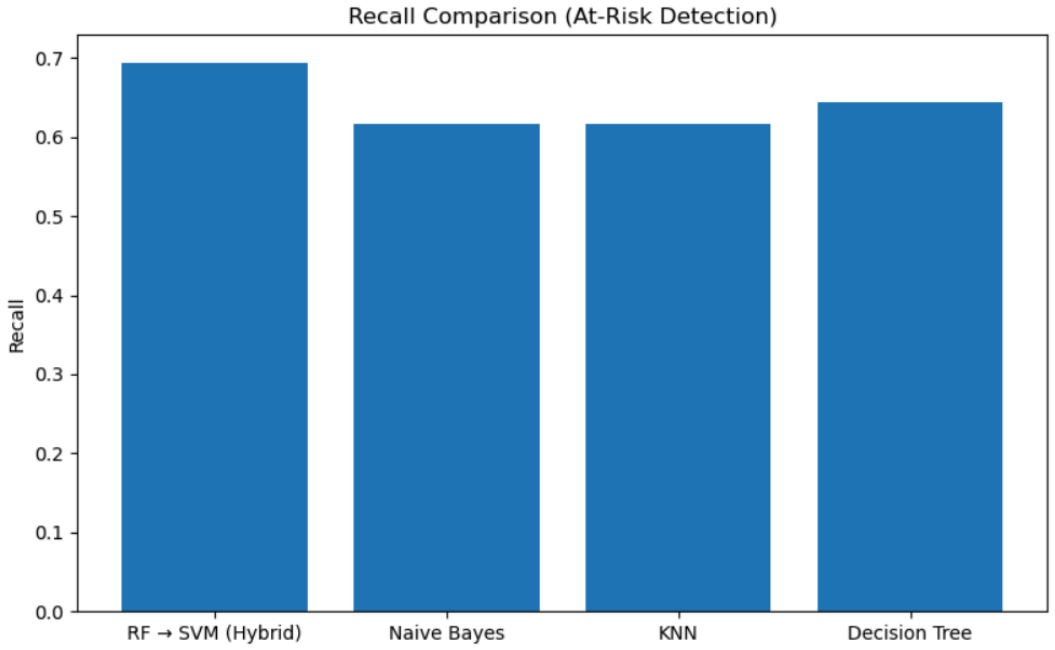

Fig 3: Recall comparison for early detection of at-risk students.

Fig 3 demonstrate that the RF → SVM hybrid model exhibits the highest recall value, indicating superior sensitivity toward identifying students who are likely to disengage or drop out. While Decision Tree and Naïve Bayes demonstrate moderate recall, their lower performance suggests a higher likelihood of false negatives. The hybrid model's improved recall can be attributed to the Random Forest's ability to isolate influential engagement features, followed by the SVM's margin-based decision boundary that enhances class separation.

Accuracy provides an overall measure of correct classifications and serves as a complementary metric to F1-score and recall. Although accuracy alone may not fully reflect performance under class imbalance, it remains useful for evaluating general predictive reliability.

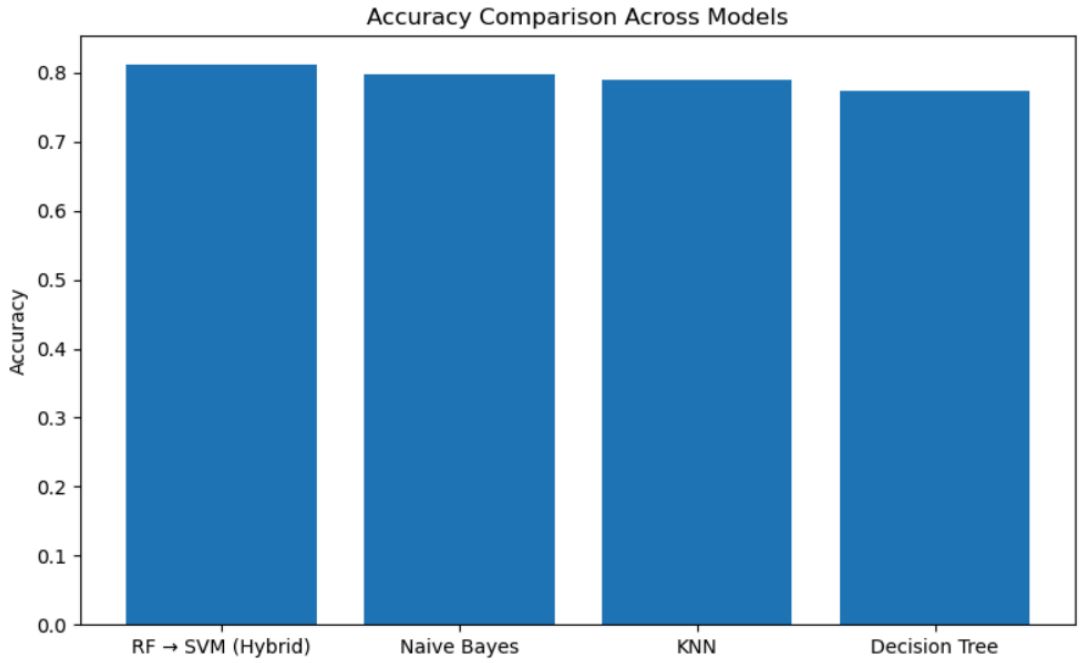

Fig 4: Accuracy comparison across baseline models and the proposed hybrid approach

Fig 4 shows that the hybrid RF → SVM model achieves the highest accuracy among the compared algorithms, confirming that the performance gains observed in F1-score and recall do not come at the cost of overall correctness. Baseline models such as KNN and Decision Tree exhibit comparatively lower accuracy, likely due to sensitivity to noise and overfitting in high-dimensional engagement data. The results indicate that the proposed hybrid framework generalizes more effectively across diverse learner profiles.

5. CONCLUSION

This study proposed a conceptually grounded two-stage machine learning framework for the early identification of at-risk students in distance learning environments. The framework integrates systematic data preparation, Random Forest–based feature selection, and Support Vector Machine classification to address key challenges such as high dimensionality, data noise, and limited interpretability. Comparative analysis using accuracy, recall, and F1-score metrics suggests that the hybrid Random Forest → SVM approach offers improved performance over baseline models, particularly in identifying at-risk learners. Emphasis on recall is especially important in educational contexts, where early detection enables timely intervention. Although the framework is not empirically implemented, it provides a robust and scalable foundation for future validation using real-world learning management system data and for supporting effective early-warning systems.

REFERENCES

1. Wang, J., & Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. *PLOS ONE, 20*.

2. Alnasyan, B., Basheri, M., & Alassafi, M.O. (2024). The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review. *Comput. Educ. Artif. Intell., 6*, 100231.

3. Al-Shabandar, R., Hussain, A.J., Keight, R., & Khan, W. (2020). Students Performance Prediction in Online Courses Using Machine Learning Algorithms. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1-7.

4. Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A.S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*, 1 - 26.

5. Kotsiantis, S.B., Pierrakeas, C.J., & Pintelas, P.E. (2004). PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES. *Applied Artificial Intelligence, 18*, 411 - 426.

6. Ahmed, E. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Appl. Comput. Intell. Soft Comput., 2024*, 4067721:1-4067721:15.

7. Al-Shabandar, R., Hussain, A.J., Keight, R., & Khan, W. (2020). Students Performance Prediction in Online Courses Using Machine Learning Algorithms. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1-7.
8. Biswas, U., & Bhattacharya, S. (2023). ML-based intelligent real-time feedback system for blended classroom. *Education and Information Technologies, 29*, 3923 - 3951.

9. Zerkouk, M., Mihoubi, M., & Chikhaoui, B. (2025). SentiDrop: A Multi Modal Machine Learning model for Predicting Dropout in Distance Learning. *ArXiv, abs/2507.10421*.

10. Holicza, B., & Kiss, A. (2023). Predicting and Comparing Students' Online and Offline Academic Performance Using Machine Learning Algorithms. *Behavioral Sciences, 13*.