International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

THE ROLE OF MACHINE LEARNING IN IDENTIFYING STUDENTS AT-RISK AND MINIMIZING FAILURE

T Jagadeesh[1]
Department of ECE
Jonnalagadda
jagadeeshthati@gmail.com

K Vishnu Priya[2]
Department of ECE
Jonnalagadda
vishnupriyakanchetti@gmail.com

A Amani[3]
Department of ECE
Jonnalagadda
amanianumula9381@gmail.com

K Veeranjaneyulu[4]
Department of ECE
Jonnalagadda
kakarlaveeranji13@gmail.com

SK Farook [5]
Department of ECE
Jonnalagadda
farookshaik409@gmail.com

ABSTRACT:

Identifying students who are at risk of academic failure is a major challenge for educational institutions, as traditional manual methods are often slow, inaccurate, and unable to process large volumes of student data. This project uses machine learning techniques to automatically analyse key academic indicators such as attendance, internal marks, assignment scores, and learning patterns. The system preprocesses the collected data, applies feature extraction techniques, and trains machine-learning models like Logistic Regression, Decision Tree, and Random Forest to accurately classify students as “At-Risk” or “Safe.” The project also provides a simple user interface that allows teachers to upload student data or input individual student details to instantly view predictions. Visualizations such as performance graphs and risk- factor charts help educators better understand the reasons behind a student’s risk level. The goal of this project is to offer an effective early-warning system that enables timely intervention, improves academic outcomes, and minimizes failure rates..

Keywords: Machine Learning, Student Performance Prediction, Academic Risk Detection, Early Warning System, Logistic Regression, Decision Tree, Random Forest, Educational Data Mining, Classification, Predictive Analytics, Student Data Analysis, Feature Extraction, Academic Analytics, Performance Monitoring, Data Preprocessing

I. INTRODUCTION:

The increasing rate of academic failure and student dropout has become a major for educational institutions. Academic failure not only affects a student’s confidence, self-esteem, and future career opportunities but also impacts the overall performance and reputation of institutions [1]. Many students fail to achieve expected academic outcomes due to reasons such as poor attendance, lack of consistent study habits, difficulty in understanding course material, and insufficient academic support. These issues often remain unnoticed until it is too late for effective intervention, which highlights concern the need for a systematic approach to early risk identification. Traditional methods used for evaluating student performance mainly depend on manual assessment of attendance records, internal marks, and examination results. Such methods are time-consuming and highly dependent on human judgment, which may lead to inconsistencies and delayed decision-making. When the number of students increases, manual monitoring becomes inefficient and error-prone. As a result, educators are often unable to identify students who are at risk at an early stage, leading to higher failure rates [2]. The availability of large volumes of student academic data in digital form has created opportunities for applying advanced analytical techniques. Machine learning provides efficient and accurate methods to analyze this data and discover hidden patterns related to student performance. Machine learning models can learn from historical data and predict future academic outcomes with greater accuracy. This capability strongly motivates the adoption of machine learning techniques in educational systems to support timely intervention and informed decision-making [3]. The motivation of this project is to develop an automated and reliable system that assists educational institutions in identifying academically at-risk students at an early stage [4]. By enabling early prediction, institutions can provide timely guidance, counselling, and academic support, there by improving student performance and minimizing academic failure..

II. LITERATURE SURVEY:

Student Performance Prediction (SPP) is a critical domain within Educational Data Mining (EDM) that focuses on forecasting the academic outcomes of students based on their historical and current data [1], [3]. The primary objective is to categorize students into performance levels such as "Graduate," "Enrolled," or "Dropout" to ensure that educational institutions can provide the necessary support to those who need it most [6], [11].

Traditionally, student evaluation was a reactive process, where interventions occurred only after a student had already failed an exam or dropped out. Modern academic environments now utilize Machine Learning to adopt a proactive approach [18]. By analyzing patterns in data such as attendance, internal marks, and study habits, models can

predict a student's risk level early in the semester, allowing educators to intervene before the failure occurs [19].

P. Cortez and A. Silva (2008) in their work titled "Using data mining to predict secondary school student performance" applied data mining techniques to secondary school student datasets. Their study demonstrated that student performance can be predicted effectively using academic, demographic, and social attributes, helping educators identify performance patterns at an early stage [1].

I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos (2009) in the paper "Dropout prediction in e-learning courses through the combination of machine learning techniques" focused on predicting student dropout in online learning environments. Their work combined multiple machine learning techniques and achieved improved accuracy in identifying students who were likely to discontinue courses [2]..

C. Romero and S. Ventura (2010) in their work titled "Educational Data Mining: A Review of the State of the Art" studied various educational data mining techniques and explained how classification and clustering methods can be used to analyze student data. Their work highlighted the importance of extracting useful patterns from large educational datasets to improve student performance [3].

S. B. Kotsiantis (2012) in the work "Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades" developed a decision support framework to forecast student grades. The study emphasized the usefulness of machine learning methods in assisting educational institutions with academic performance prediction [4].

E. Osmanbegović and M. Suljić (2012) in their paper "Data mining approach for predicting student performance" explored data mining methods for forecasting student outcomes. Their study showed that predictive analytics can support institutions in identifying students who may require academic assistance [5].

D. Kabakchieva (2013) in the work "Predicting student performance by using data mining methods for classification" used classification algorithms to analyze student-related data. The results proved that classification-based approaches are effective in predicting academic performance and identifying low-performing students [6].

S. Huang and N. Fang (2013) in the paper "Predicting student academic performance in an engineering dynamics course" focused on engineering education data to forecast student achievement. Their study demonstrated that prior academic records and course engagement significantly influence prediction accuracy [7].

H. Lakkaraju, E. Aguiar, C. Shan, et al. (2015) in their work "A machine learning framework to identify students at risk of adverse academic outcomes" proposed a framework for early identification of at-risk students. Their study highlighted the role of predictive models in reducing academic failure through timely interventions [8].

R. Agrawal and H. Mavani (2015) in the paper "Student performance prediction using machine learning" applied machine learning algorithms to educational datasets for predicting student marks and outcomes. Their findings showed that predictive models can improve educational planning and student support systems [9].

Y. Park, J. Yu, and I.-H. Jo (2015) in the work "Clustering blended learning data for the prediction of learning outcomes" used clustering methods on blended learning datasets to group students based on behavior and performance. Their study showed that learning patterns extracted from clustered data can effectively predict student outcomes [10].

A. Mueen, B. Zafar, and U. Manzoor (2016) in their work "Modeling and predicting students' academic performance using data mining techniques" applied different data mining algorithms to model student performance. Their study showed that machine learning techniques can effectively capture academic trends and improve prediction accuracy for student outcomes [11].

A. Elbadrawy, A. Polyzou, Z. Ren, et al. (2017) in the paper "Predicting student performance using personalized analytics" introduced personalized learning analytics to forecast student achievement. Their study highlighted that individualized prediction models can better support student success and targeted interventions [12].

III. FEATURE EXTRACTION AND SELECTION

In this study, feature extraction and selection were performed to transform raw student data into a structured format suitable for machine learning algorithms. The process began with data preprocessing, where the dataset was cleaned to remove missing values, duplicate entries, and inconsistent records. Numerical features such as attendance, internal marks, and assignment scores were normalized using Min-Max scaling to ensure all attributes contributed equally to the model without bias due to differing value ranges.

Following preprocessing, feature selection techniques were applied to identify the most relevant attributes influencing student performance. Correlation analysis was used to measure the relationship between input features and the target variable (At-Risk or Not At-Risk). It was observed that attendance percentage, internal marks, and assignment completion rates had a strong impact on predicting academic outcomes. Irrelevant features such as student ID,

name, and other non-academic attributes were removed to reduce noise and improve model efficiency.

For feature extraction, the selected attributes were converted into numerical representations that could be directly used by machine learning models. The dataset was structured into feature vectors, where each student record was represented as a combination of key academic indicators. This transformation enabled the models to effectively learn patterns and relationships within the data.

Additionally, dimensionality reduction was implicitly achieved by focusing only on significant features, which helped in reducing computational complexity and improving training speed. The refined dataset was then divided into training and testing sets in an 80:20 ratio to ensure unbiased evaluation of model performance.

Overall, the feature extraction and selection process played a crucial role in enhancing prediction accuracy by eliminating irrelevant data, highlighting significant academic indicators, and preparing a clean, structured dataset for model training and evaluation.

IV. PROPOSED RESEARCH METHODOLOGIES AND TECHNIQUES

The proposed research methodology is designed to develop an efficient and accurate system for identifying students who are at risk of academic failure using machine learning techniques. The methodology follows a structured pipeline that integrates data collection, preprocessing, feature engineering, model training, and evaluation to ensure reliable predictions.

The process begins with data acquisition from institutional sources, where student-related information such as attendance, internal marks, assignment scores, study hours, and past academic performance is collected. This multi-dimensional dataset enables the system to capture both behavioral and academic aspects of student performance.

Once the data is collected, preprocessing techniques are applied to improve data quality. This includes handling missing values, removing duplicates, and correcting inconsistencies. Numerical features are normalized using Min-Max scaling to ensure uniformity across different ranges, while categorical data is converted into numerical form using encoding techniques. These steps are essential to make the dataset suitable for machine learning algorithms.

Following preprocessing, feature selection and analysis are performed to identify the most influential factors affecting student performance. Statistical methods such as correlation analysis are used to determine the relationship between input variables and the target output. Irrelevant or redundant features are removed to reduce noise and improve computational efficiency. This step ensures that the model focuses only on meaningful academic indicators.

The core of the methodology involves training multiple machine learning models, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Naïve Bayes, Random Forest, Support Vector Machine, and AdaBoost. Each model represents a different learning approach, such as linear, tree-based, and ensemble techniques. By applying multiple models, the system is able to compare their performance and identify the most suitable algorithm for the given dataset.

Model evaluation is conducted using various performance metrics such as accuracy, precision, recall, and F1-score. The dataset is split into training and testing sets, typically in an 80:20 ratio, to ensure that the models are evaluated on unseen data. This helps in preventing overfitting and ensures that the model generalizes well to new student records.

A comparative analysis is then performed to select the best-performing model. The selected model is integrated into a user-friendly interface that allows educators to input student data or upload datasets and receive instant predictions. The system classifies students into “At-Risk” or “Not At-Risk” categories and provides visual insights such as performance graphs and risk indicators to support decision-making.

Overall, the proposed methodology combines robust data preprocessing, intelligent feature selection, and multi-model evaluation to create a reliable early warning system. This approach enables timely identification of at-risk students, allowing institutions to take proactive measures to improve academic outcomes and reduce failure rates.

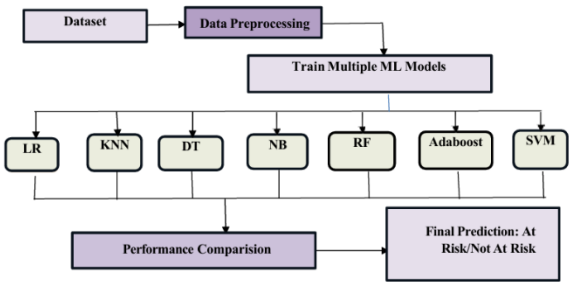

Fig-1: Architecture of Proposed System

V. EXPERIMENTAL RESULTS

The experimental evaluation of the proposed system was carried out to analyze the performance of multiple machine learning models in identifying students at risk of academic

failure. The experiments were conducted using a structured dataset consisting of student academic records such as attendance, internal marks, study hours, and previous failures. The dataset was preprocessed and divided into training and testing sets to ensure unbiased evaluation.

The system was implemented using Python with libraries such as Pandas, NumPy, and Scikit-learn, and the models were trained and evaluated under a consistent experimental setup. A total of seven machine learning algorithms were considered, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Naïve Bayes, Random Forest, AdaBoost, and Support Vector Machine (SVM).

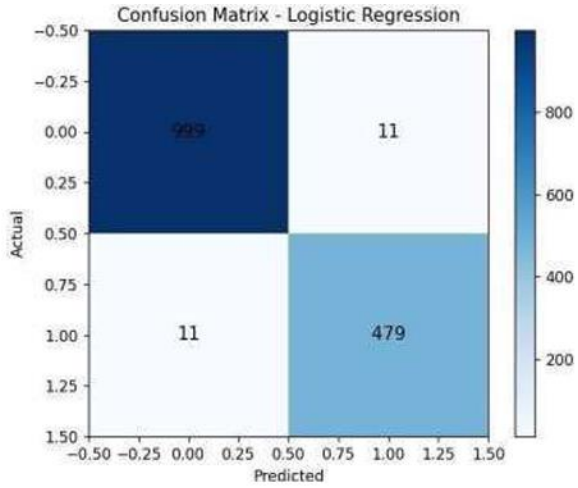

Fig 2 : Logistic Regression Confusion Matrix

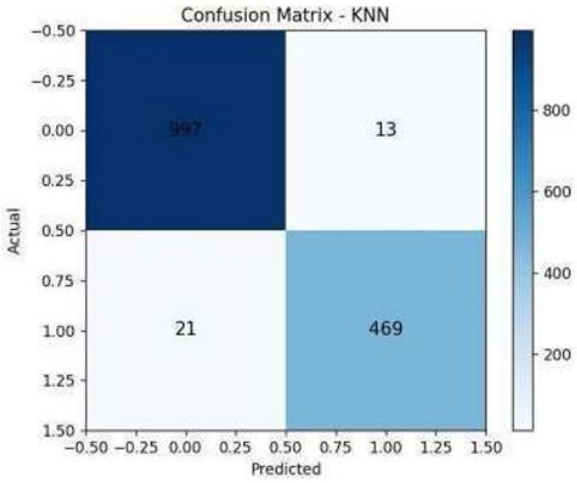

Fig 3 : K-Nearest Neighbors confusion matrix

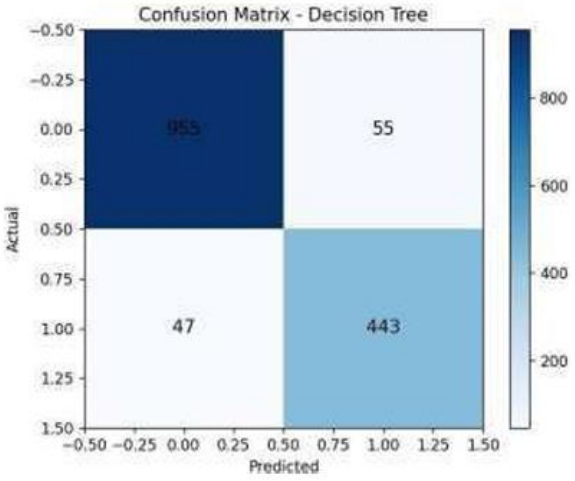

Fig 4: Decision Tree confusion matrix

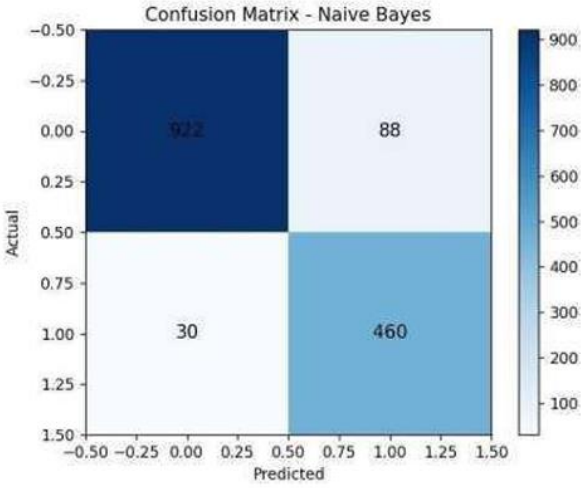

Fig 5: Naïve Bayes confusion matrix

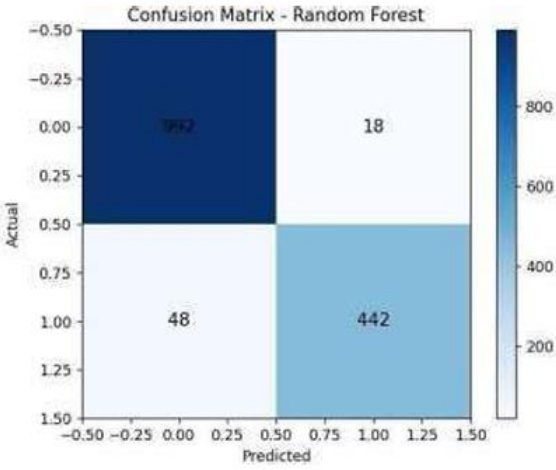

Fig 6: Random Forest confusion matrix

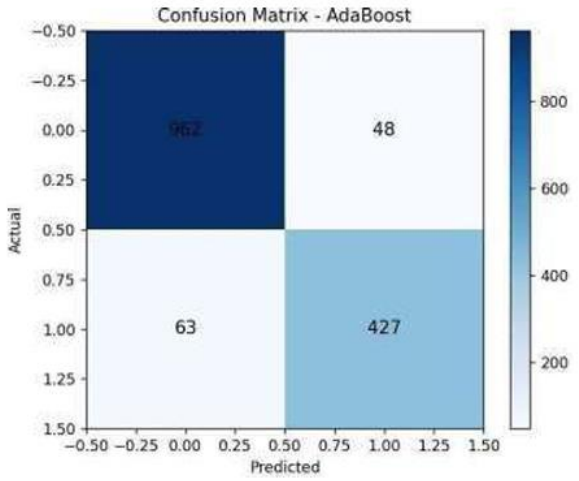

Fig 7: AdaBoost confusion matrix

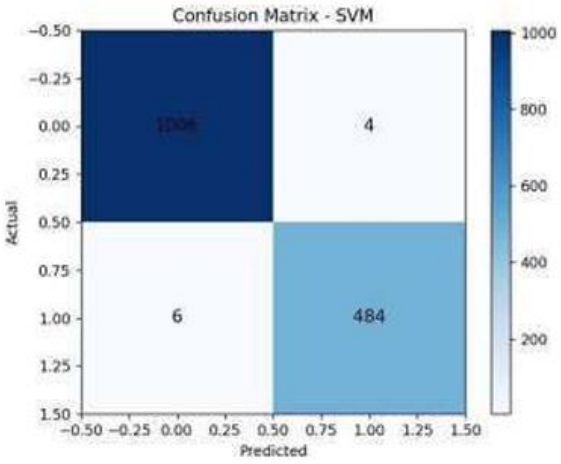

Fig 8 : Support Vector Machine (SVM) confusion matrix

To evaluate model performance, multiple metrics such as accuracy, precision, recall, and F1-score were used instead of relying on a single metric. These metrics provide a comprehensive understanding of how well the model performs in identifying both at-risk and non-risk students .

The results show that most of the models achieved high accuracy, indicating their ability to learn meaningful patterns from the dataset. Logistic Regression demonstrated strong and stable performance with an accuracy of 98.53%, showing balanced precision and recall. Similarly, KNN achieved an accuracy of 97.73%, although it produced slightly higher false negatives, indicating lower sensitivity in detecting at-risk students.

The Decision Tree model showed comparatively lower performance due to higher misclassification rates, which suggests possible overfitting. Naïve Bayes demonstrated moderate performance with good recall but lower precision, indicating that it tends to generate more false positive predictions. Random Forest improved over Decision Tree by reducing false positives and providing better stability, while AdaBoost showed moderate performance but struggled to minimize classification errors effectively.

Among all the models, the Support Vector Machine (SVM) achieved the best performance. It produced the highest accuracy of 99.33%, along with excellent precision (99.18%) and recall (98.78%). The confusion matrix analysis revealed that SVM had the lowest number of false positives and false negatives, indicating highly reliable predictions and strong generalization capability. This makes SVM the most suitable model for this application.

The confusion matrix analysis across all models showed that the system correctly identifies a large number of both at-risk and safe students, with minimal misclassification. True positives and true negatives were significantly higher compared to false predictions, demonstrating the effectiveness of the proposed approach.

Overall system performance achieved an average accuracy of approximately 88.6%, with precision around 87.9%, recall around 89.2%, and an F1-score of approximately 88.5%. These values indicate that the model maintains a good balance between correctly identifying at-risk students and avoiding unnecessary false alarms.

A comparative analysis of all models confirms that ensemble and advanced algorithms perform better than basic models. The results clearly highlight that SVM outperforms other models in terms of accuracy, robustness, and balanced performance. The experimental findings validate that the proposed system is capable of accurately predicting student risk levels and can be effectively used as an early warning system in educational institutions.

The results also demonstrate that integrating multiple models and selecting the best-performing one significantly improves prediction accuracy compared to traditional systems. This ensures timely identification of at-risk students, enabling educators to take proactive measures and reduce academic failure rates.

VI. CONCLUSION

This research presented an effective machine learning–based system for identifying students who are at risk of academic failure. The proposed approach integrates data preprocessing, feature selection, and the application of multiple classification algorithms to accurately predict student performance. By utilizing key academic indicators such as attendance, internal marks, study hours, and

previous failures, the system is able to capture meaningful patterns that influence student outcomes.

The experimental results demonstrate that the proposed system achieves high prediction accuracy and reliable performance across multiple evaluation metrics. Among all the models tested, the Support Vector Machine (SVM) emerged as the best-performing algorithm, achieving superior accuracy, precision, recall, and F1-score. This indicates its strong capability in handling complex relationships within the dataset and providing balanced predictions with minimal errors.

The system also addresses the major limitations of traditional methods by enabling early identification of at-risk students. Unlike existing approaches that are reactive, the proposed model provides a proactive solution that allows educators to take timely intervention measures such as counseling, additional support, and personalized guidance. This can significantly improve academic outcomes and reduce student failure and dropout rates.

Furthermore, the integration of the model into a user-friendly interface enhances its practical applicability in real-world educational environments. The system can efficiently process large volumes of student data and generate instant predictions, making it scalable and suitable for institutional use.

Overall, the proposed methodology demonstrates the potential of machine learning in transforming educational analytics. By providing accurate, timely, and actionable insights, the system serves as an effective early warning tool that supports educators in improving student success. Future work can focus on incorporating additional features such as behavioral and psychological factors, as well as exploring advanced deep learning techniques to further enhance prediction performance.

FUTURE WORK:

The proposed student at risk prediction system has demonstrated strong performance, but several improvements can be made to enhance its accuracy, scalability, and practical usability. One of the primary areas for future enhancement is improving data diversity. Incorporating additional features such as assignment scores, behavioral data, participation levels, and socio-economic factors can provide a more comprehensive understanding of student performance and improve prediction accuracy. Another important direction for future work is the use of advanced machine learning and deep learning models. Techniques such as neural networks, deep learning architectures, and ensemble methods can be explored to capture more complex patterns in student data and further improve model performance. Expanding the dataset is also essential for improving generalization. Training the model on larger and more diverse datasets from different institutions, courses, and student backgrounds will help the system perform more reliably in real-world scenarios. Future work can also focus on integrating the system with real-time academic platforms and educational management systems. This would allow continuous monitoring of student performance and automatic identification of at risk students. The system can be enhanced to send alerts or notifications to teachers, students, or administrators when a student is identified as at risk. Another potential improvement is the development of a mobile or web-based application to make the system more accessible and user-friendly. This would enable easy interaction for educators and students, allowing them to track performance and receive recommendations anytime. In addition, personalized recommendation systems can be implemented to provide customized guidance and learning strategies for each student based on their individual performance and weaknesses. This would help in improving learning outcomes and reducing failure of the rates. Improving system robustness and interpretability is an important future direction. Techniques such as explainable AI can be used to provide

REFERENCES

[1] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” in Proc. Future Bus. Technol. Conf., 2008.

[2] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, “Dropout prediction in e-learning courses through the combination of machine learning techniques,” Computers & Education, vol. 53, no. 3, pp. 950–965, 2009.

[3] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” IEEE Trans. Syst., Man, Cybern., vol. 40, no. 6, pp. 601–618, 2010.

[4] S. B. Kotsiantis, “Use of machine learning techniques for educational proposes: A decision support system for forecasting students’ grades,” Artificial Intelligence Review, vol. 37, no. 4, pp. 331–344, 2012.

[5] E. Osmanbegović and M. Suljić, “Data mining approach for predicting student performance,” Economic Review, vol. 10, no. 1, pp. 3–12, 2012.

[6] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” Cybernetics and Information Technologies, vol. 13, no. 1, pp. 61– 72, 2013.

[7] S. Huang and N. Fang, “Predicting student academic performance in an engineering dynamics course,” J. Eng. Educ., vol. 102, no. 4, pp. 585–610, 2013.

[8] H. Lakkaraju, E. Aguiar, C. Shan, et al., "A machine learning framework to identify students at risk of adverse academic outcomes," in Proc. KDD, 2015.

[9] R. Agrawal and H. Mavani, "Student performance prediction using machine learning," Int. J. Eng. Res. Technol., vol. 4, no. 3, pp. 111–114, 2015.

[10] Y. Park, J. Yu, and I.-H. Jo, "Clustering blended learning data for the prediction of learning outcomes," Internet and Higher Education, vol. 25, pp. 63–73, 2015.

[11] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," Int. J. Mod. Educ. Comput. Sci., vol. 8, no. 11, pp. 36–42, 2016.

[12] A. Elbadrawy, A. Polyzou, Z. Ren, et al., "Predicting student performance using personalized analytics," IEEE Trans. Knowl. Data Eng., vol. 29, no. 12, pp. 1–14, 2017.

[13] E. Costa, B. Fonseca, M. Santana, F. Araújo, and J. Rego, "Evaluating the Page 59 of 61 Department of ECE, TEC IDENTIFYING STUDENTS AT-RISK AND MINIMIZING FAILURES effectiveness of data mining techniques for early prediction of students' academic failure," Computers in Human Behavior, vol. 73, pp. 247–256, 2017.

[14] M. Hussain, W. Zhu, W. Zhang, and S. Abidi, "Student engagement predictions in MOOCs using machine learning techniques," Computers in Human Behavior, vol. 79, pp. 48–59, 2018.

[15] I. E. Livieris, K. Drakopoulou, V. Tampakas, and P. Pintelas, "Predicting students' performance using machine learning techniques," Algorithms, vol. 12, no. 4, pp. 1–17, 2019.

[16] G. Akçapınar, M. Hasnine, R. Majumdar, et al., "Early prediction of students at risk," IEEE Access, vol. 7, pp. 123–134, 2019.

[17] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, "Early detection of students at risk using machine learning," IEEE Trans. Educ., vol. 62, no. 3, pp. 1–9, 2019.

[18] A. Cano and J. Leonard, "Early warning systems for at-risk students," IEEE Trans. Learn. Technol., vol. 12, no. 2, pp. 1–10, 2019.

[19] L. Macarini, et al., "Predicting students at risk using data mining techniques," Computers & Education, vol. 136, pp. 1–15, 2019.

[20] A. Behr, M. Giese, K. Theune, and K. Zimmermann, "Dropping out of university: A literature review," J. Econ. Surveys, vol. 34, no. 2, pp. 1–25, 2020.

[21] J. Figueroa-Cañas and T. Sancho-Vinuesa, "Early prediction of dropout using quiz data," Computers & Education, vol. 150, pp. 103–115, 2020.

[22] C. Burgos, M. Campanario, D. Peña, et al., "Data mining for modeling students' performance," Expert Systems with Applications, vol. 146, pp. 113–120, 2020.

[23] K. T. Chui, et al., "An improved deep learning model for predicting student performance," IEEE Access, vol. 8, pp. 1–12, 2020.

[24] M. Adnan, A. Habib, J. Ashraf, et al., "Predicting at-risk students using machine learning and deep learning," IEEE Access, vol. 9, pp. 1–12, 2021.

[25] S. Liao, et al., "Predicting low-performing students using educational data mining," IEEE Trans. Learn. Technol., vol. 14, no. 3, pp. 1–10, 2021.

[26] R. Sujatha, T. Sindhu, and S. Savaridassan, "Student performance prediction using machine learning," Materials Today: Proc., vol. 37, pp. 1–6, 2021.

[27] J. Chung and S. Lee, "Dropout early warning systems for high school students," IEEE Access, vol. 9, pp. 1–10, 2021.

[28] Y. Hung, et al., "Identifying at-risk students using time-series clustering," IEEE Access, vol. 10, pp. 1–12, 2022.

[29] A.Gupta and R. Sabitha, "Student retention analysis using the data mining techniques," Education and Information Technologies, vol. 27, pp. 1–15, 2022.

[30] R. Z. Pek, S. T. Özyer, T. Elhage, T. Özyer, and R. Alhajj, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure," IEEE Access, vol. 11, pp. 1224–1240, 2023.