Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hybrid Deep Learning Approach for Education Data Mining and Improvement of Education Practices

Shrishail Patil[1], Dr. Pratap Singh Patwal[2], Dr. Vinod S. Wadne[3]
[1]*Research Scholar (Computer Science Engineering), Nirwan University Jaipur, Rajasthan, India*
Shri.patil11@gmail.com
[2]*School of Engineering & Technology, Nirwan University Jaipur, Rajasthan, India*
Pratappatwal@gmail.com
[3]*HOD of Computer Engineering, JSPM's Imperial College of Engineering & Research, Pune, Maharashtra, India*
vinods1111@gmail.com

ARTICLE INFO

Received: 01 Nov 2024

Revised: 25 Dec 2024

Accepted: 08 Jan 2025

ABSTRACT

The exponential growth of educational data necessitates innovative approaches to mining and utilizing this information to enhance educational practices. This study proposes a hybrid deep learning framework for educational data mining (EDM) that integrates various data sources, advanced feature engineering techniques, and state-of-the-art classification algorithms to improve learning outcomes and institutional decision-making processes. The research utilizes a diverse dataset comprising EDM applications, real-time educational data, and synthetic student data to develop robust models. Feature engineering is conducted using a hybrid approach that combines TF-IDF, N-gram, bigram relational models, autoencoders, and density-based techniques, aiming to maximize data representation and reduce dimensionality. The classification phase incorporates an array of traditional and deep learning methods, including Naïve Bayes (NB), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest (RF), AdaBoost, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and a novel hybrid RNN-SVM model. The proposed hybrid RNN-SVM classifier demonstrates superior accuracy and scalability by leveraging RNN's sequential learning capabilities and SVM's margin-based classification efficiency. Additionally, a recommendation module is designed to provide actionable insights, including class improvement strategies and industry-oriented suggestions, thus bridging the gap between academic performance and professional readiness. The hybrid deep learning framework not only enhances predictive accuracy but also facilitates informed decision-making for educators and policymakers. Experimental results validate the framework's efficacy in mining meaningful patterns from complex educational datasets and optimizing learning strategies. This research highlights the transformative potential of hybrid deep learning in advancing the field of EDM and fostering improved educational practices.

Keywords: EDM, Hybrid Deep Learning, Educational Practices, Student Performance Prediction, Learning Analytics, Personalized Learning, Educational Recommendation Systems.

Introduction

The rapid advancement of technology has resulted in an exponential increase in the volume of educational data generated across various institutions worldwide. This surge in data presents significant opportunities for improving educational practices, but it also poses challenges related to efficiently mining and utilizing this information. To address these challenges, this study introduces a hybrid deep learning framework specifically designed for Educational Data Mining (EDM). By combining diverse data sources, advanced feature engineering techniques, and cutting-edge classification algorithms, this framework aims to optimize learning outcomes and support evidence-based decision-making within educational institutions. The proposed framework employs a diverse dataset comprising real-time educational data, applications of EDM, and synthetic student data, ensuring robust model development and generalizability. Central to the framework is its hybrid approach to feature engineering, which integrates techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), N-grams, bigram relational models, autoencoders, and density-based methods. These techniques work collaboratively to maximize the

representation of complex educational data while simultaneously reducing its dimensionality, enabling more efficient processing and analysis. This feature engineering process ensures that the models can uncover hidden patterns and insights that are critical for improving educational practices.

The classification phase of the framework leverages a variety of machine learning and deep learning algorithms, ranging from traditional methods like Naïve Bayes (NB) and Support Vector Machines (SVM) to more sophisticated approaches such as Artificial Neural Networks (ANN), Random Forest (RF), and AdaBoost. Additionally, advanced models like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and a novel hybrid RNN-SVM classifier are implemented to address the unique challenges of educational data. The hybrid RNN-SVM model stands out due to its ability to combine the sequential learning capabilities of RNNs with the margin-based classification efficiency of SVMs, resulting in superior accuracy, scalability, and performance. This innovative classifier demonstrates significant potential for advancing the field of EDM. Beyond classification, the framework includes a recommendation module designed to translate data-driven insights into actionable strategies. This module offers practical suggestions for improving classroom teaching methods, enhancing academic performance, and aligning educational curricula with industry requirements. By bridging the gap between academic achievement and professional readiness, the framework empowers educators, administrators, and policymakers to make informed decisions that positively impact both individual learners and broader institutional goals.

The experimental results presented in this research validate the effectiveness of the hybrid deep learning framework in mining meaningful patterns from complex and diverse educational datasets. These findings demonstrate the transformative potential of leveraging hybrid models to address the multifaceted challenges faced in education today. By enabling more accurate predictions and facilitating better decision-making processes, this framework paves the way for innovative applications in EDM. It highlights how technology can be harnessed to not only enhance educational practices but also shape a more informed and adaptive educational system for the future. This paper underscores the importance of integrating advanced deep learning techniques into the field of EDM. The proposed hybrid framework sets a new standard for predictive accuracy and actionable insights, fostering a data-driven approach to education that can drive meaningful change across the sector.

Literature Survey

According to Usman Ali et al. [1] in 2019, huge numbers of records in educational data sets can make it difficult to produce data of high quality. Many scholars in the field of education are now analyzing data utilizing the DM technique. However, instead doing feature extraction on data, a number of research papers concentrate on choosing the best learning method. As a result, completing categorization on the dataset demands a significant amount of processing time due to its high computational effort. This document provides an overview of the feature selection techniques used for the evaluation of data characteristics. In order to improve the quality of the students' data collection, the suggested hybrid method integrates feature extraction with wrapper-based methods.

According to Gabriella Casalino et al. [2] in 2020, Virtual Learning Environments (VLEs) are Web-based systems that offer instructional materials as well as study aids. Daily logs of student engagement with VLEs are gathered, hence computerized solutions are expected to handle and evaluate such enormous amounts of data. The insights derived from educational data can be utilized by pupils, instructors, administrators, and generally all stakeholders participating in the VLEs' learning activities. Valuable information can be found utilizing ML approach. Historically, static ML techniques have been used to analyze educational data sets. Nevertheless, because they are by their very nature non-stationary, educational dataset are best handled as data streams. The findings of a classification research in which the RF method was employed to create a model for forecasting the passing or failing of students' tests were presented in this publication. The strongest discriminating qualities for the prediction task are also found via a feature importance assessment. On the Open University Learning Analytics Database (OULAD), tests were conducted to demonstrate the accuracy of adaptive RF in producing classification techniques from changing data sets.

A paradigm with standard recommendations for a productive and effective data collection technique for an EDM research is proposed by Bhanuka Mahanama et al. in 2018 [3]. By assuring an effective and thorough data gathering, the report illustrates prospective difficulties and approaches to address some significant obstacles. These approaches also offer helpful validation methods to guarantee the data's internal consistency and little duplication. Ultimately, the research identifies significant concerns with academic data gathering for EDM, along with potential approaches to solve these problems. The paper suggests a model for categorizing the data needs for an effective DM procedure in order to handle the major problem of the recognition of pertinent measures. The suggested technique offers evidence of its effectiveness, thoroughness, and thoroughness for data collecting, despite the difficulty in developing an ideal

technique. These truths are emphasized by surveys that are internally consistent and relatively duplicated, as shown by the studies. The results of this study can also be applied to the creation of a sophisticated system for gathering educational data. Additionally, the research can be seen as the first data mining-based study conducted for the Sri Lankan educational system, and the recommendations can offer insightful information and discoveries for future developments.

According to Michal Munk et al. [4] Preprocessing log files for educational data is a time-consuming step in the knowledge discovery method. Data cleansing, user recognition, session detection, and path fulfilment phase are all parts of this process. The research makes an effort to pinpoint the steps that must occur when preparing educational information in order to apply LA techniques afterwards. The paper attempts to give a response of which of these pre - processing stage has a massive effect on discovered knowledge in overall, as well as in the interpretation of amount and quality of found sequence patterns, because the sequential methods assessment is thought to be appropriate for computation of discovered knowledge. As a result, different levels of data preparation were needed to enable log files using a variety of information pre-processing approaches for session recognition and path fulfilment. The findings demonstrated that the efficiency of the retrieved sequence rules was significantly impacted by the session identification utilizing the reference length derived from the sitemap. Only the number of retrieved sequence rules was significantly impacted by the path completion method. The discovered results, when combined with the findings of earlier systematic research in educational data pre-processing, can enhance the mechanisation of the educational data pre - processing stage and also create LA tools that are appropriate for various stakeholder groups involved in EDM research tasks.

Data mining is currently applied in a variety of fields, most notably in student educational and learning analytics, according to Akansha Mishra et al. report's from 2017 [5]. Discovering hidden knowledge in information individually is very difficult and time-consuming. Clustering will be applied in the article to enhance EDM. Researchers used the data from 84 undergraduate students and divided the pupils based on the final grades they earned in the course in order to enhance effectiveness and the unambiguousness of the models that were produced. The outcome demonstrates that a particular model's clarity and unambiguity are both significantly higher than those of a general framework.

According to John Jacob et al. [6] the analysis and examination of data from academic databases is the focus of the burgeoning learning science field of EDM. One may investigate, anticipate, and enhance a student's achievement through the analysis of these vast datasets utilizing a variety of DM techniques. This article reports on a study on several EDM approaches and how they might be applied to the advantage of all the educational system's stakeholders. To determine if a change in one variable causes a change in the other, covariance is used. DTs are employed in this study to forecast student productivity by implementing potential outcomes. When building a model with a dependent variable and numerous independent factors, regression testing is carried; if the analysis is successful, the value of the dependent variable is calculated using the numbers of the independent factors. In order to arrange the objects under examination and to analyze the job descriptions that would be best for every pupil, clustering identifies groupings of items so that they are more similar to one another than to items in other clusters.

According to El Harrak Othman et al. [7] DM in education is becoming a more hot research topic as a result of the increased interest in both fields. To extract hidden information from the data sources, several DM methods, including categorization and clustering, can be used. Online video mining is the process of extracting information from the World Wide Web by utilizing DM tools. Conventional image processing and a metadata-based technique are the two methods used for web video extraction. It is specifically focused on EDM and MOOCs, which are a brand-new form of online learning. It offers an approach for mining MOOC films that makes advantage of metadata as the primary source of new information.

The study of EDM by Lixia Ji et al. [8] has advanced quickly in 2020. Furthermore, the majority of studies concentrate on problems with the data sources and downplay the significance of data pre-processing as well as DM methods. This research examined EDM with an emphasis on techniques for educational large DM. In the beginning, it examined the pertinent EDM components and introduced big data technique in accordance with the demands of academic data applications. The general educational large DM methods and their uses were then covered, followed by a discussion of the future directions for these techniques.

According to Ashish Dutt et al. [9] currently, educational institutions gather and keep enormous amounts of data about students' enrolment, attendance, and exam outcomes. This data can be mined to produce interesting data to assist to those who manage it. The explosive increase of educational data indicates that more complex techniques are

needed to distil big amounts of information. Due to this problem, the area of EDM has emerged. Conventional DM techniques may have a defined goal and functionality, thus they cannot be used to solve educational issues directly. This suggests that before applying suitable DM techniques to the issues, a pre-processing procedure must first be implemented. Grouping is one such pre-processing tool used in EDM. The applicability of various DM algorithms to educational qualities has been the topic of several researches on EDM. As a result, this study presents a thorough analysis of the clustering method's applicability and usefulness in the framework of EDM over the course of more than 30 years (1983–2016). On the basis of the literature assessment, future discoveries are presented, and opportunities for additional research are noted.

Using several EDM clustering techniques, J. L. C. Ramos et al. [10] described the gathering of information from the Moodle repository of a novice distance learning class at a Federal University in 2016. As per the connection and performance features of the various student societies, it has performed grouping using hierarchical as well as non-hierarchical approaches. In the assessment, it was probable to see how the groups acquired similarity between the outcomes of every technique utilised, supporting the knowledge gained from the clustering and illuminating how little impact the method choice in this research had on the acquired knowledge from connections and students' achievement in the course.

According to Manas Chaturvedi et al. [11] Due to its contribution in determining student failure or the learner at risk of failing a subject by applying numerous potent DM methods, EDM is an emerging area in the DM field. This article covers a collection of earlier research studies in the area of electronic dance music. It provides a basic overview of the science of DM, EDM, various tools utilized in the process, etc. Including the various steps the data set must go through before being provided to the classification model for the purpose of producing findings, the objectives of EDM have also been described. The various techniques used in the area of EDM are briefly described, along with some of their fundamental characteristics.

A special session on the application of computational methods for the analysis of educational data is proposed by Camilo Vieira et al. for 2019 [12]. Because it offers distinctive ways to describe information and comprehend complicated processes, computation has influenced all academic fields. To better comprehend educational phenomena, fields like LA and EDM have evolved in education. Three alternative strategies for using analytical tools to examine qualitative educational data will be covered in this special session. Following the lecture, the participants are expected to put these techniques into practice by using R language while considering how they may apply these techniques to their own situations.

This work in the application of computational techniques for academic research will need multidisciplinary cooperation between computer scientists and educationalists to fully connect these approaches. Recently, a variety of topics have emerged, including specific conferences and publications, as well as special issues in significant publications in education. These fields include EDM, LA and ML for education. Furthermore, a survey conducted for this paper revealed that computational specialists occasionally concentrate on complex techniques and visualization tactics with little relation to pedagogical literature. Similarly, educationists get their study conclusions from theories of teaching, but they rarely consult specialized material on computational tools or graphics. Future systems will encourage this fusion of the two concepts in order to use technology effectively for educational phenomena.

Research Methodology

Data Sources: In the hybrid deep learning approach for Education Data Mining (EDM), the quality and diversity of data sources play a critical role. The data originates from three primary sources: real-time educational data, synthetic educational data, and EDM applications. Real-time educational data includes live student interactions with Learning Management Systems (LMS), online testing platforms, and classroom sensors that track attendance, engagement, and progress. This data provides rich, dynamic information, capturing student behaviors and learning patterns. On the other hand, synthetic educational student data is generated to fill gaps in real-world data, ensuring scalability and diversity. By using simulations and data generation techniques, researchers can model realistic educational environments, assess rare events, and overcome privacy challenges. For example, synthetically generated student performance datasets may simulate various learning scenarios, capturing edge cases or outliers.

Finally, EDM applications serve as robust repositories of structured and unstructured data. These include platforms like Intelligent Tutoring Systems (ITS), adaptive learning platforms, and educational data portals. Such applications generate valuable metadata, including error logs, feedback reports, and performance summaries, crucial for understanding student progress and the effectiveness of teaching methodologies. Integrating these data sources provides a rich foundation for hybrid deep learning models, enabling comprehensive insights into educational

practices. The diversity of data helps balance model training, reducing overfitting and enhancing generalization. However, challenges like data cleaning, ethical use, and integration require meticulous preprocessing and adherence to data privacy standards.

Feature Engineering : Feature engineering is pivotal in the hybrid deep learning approach for EDM. It involves transforming raw data into meaningful features that enhance the predictive power of models. Among the most effective techniques, TF-IDF (Term Frequency-Inverse Document Frequency) is widely used for extracting features from textual data, such as student feedback, teacher notes, or educational forum discussions. This method emphasizes unique, contextually significant terms, enabling better analysis of textual content. N-gram and bigram relational techniques further contribute by identifying patterns and dependencies within textual data. For instance, analyzing sequences of words or phrases in student essays can uncover underlying learning challenges or behavioral trends. These techniques are especially powerful in capturing linguistic nuances in communication.

Autoencoders, an unsupervised feature extraction method, automatically learn compressed representations of complex educational data. By encoding and decoding features, autoencoders help reduce dimensionality, retaining only the most critical aspects of the data. Density-based techniques are employed to detect patterns or anomalies in datasets, such as clustering students based on similar learning behaviors or detecting outliers in performance trends. Finally, the hybrid approach combines all these methods to extract meaningful features holistically. For instance, using TF-IDF and autoencoders in tandem can extract both linguistic and structural features, while N-grams and density-based methods provide insights into relationships and anomalies. The hybridization ensures that all relevant aspects of educational data are captured, significantly enhancing the performance of downstream models.

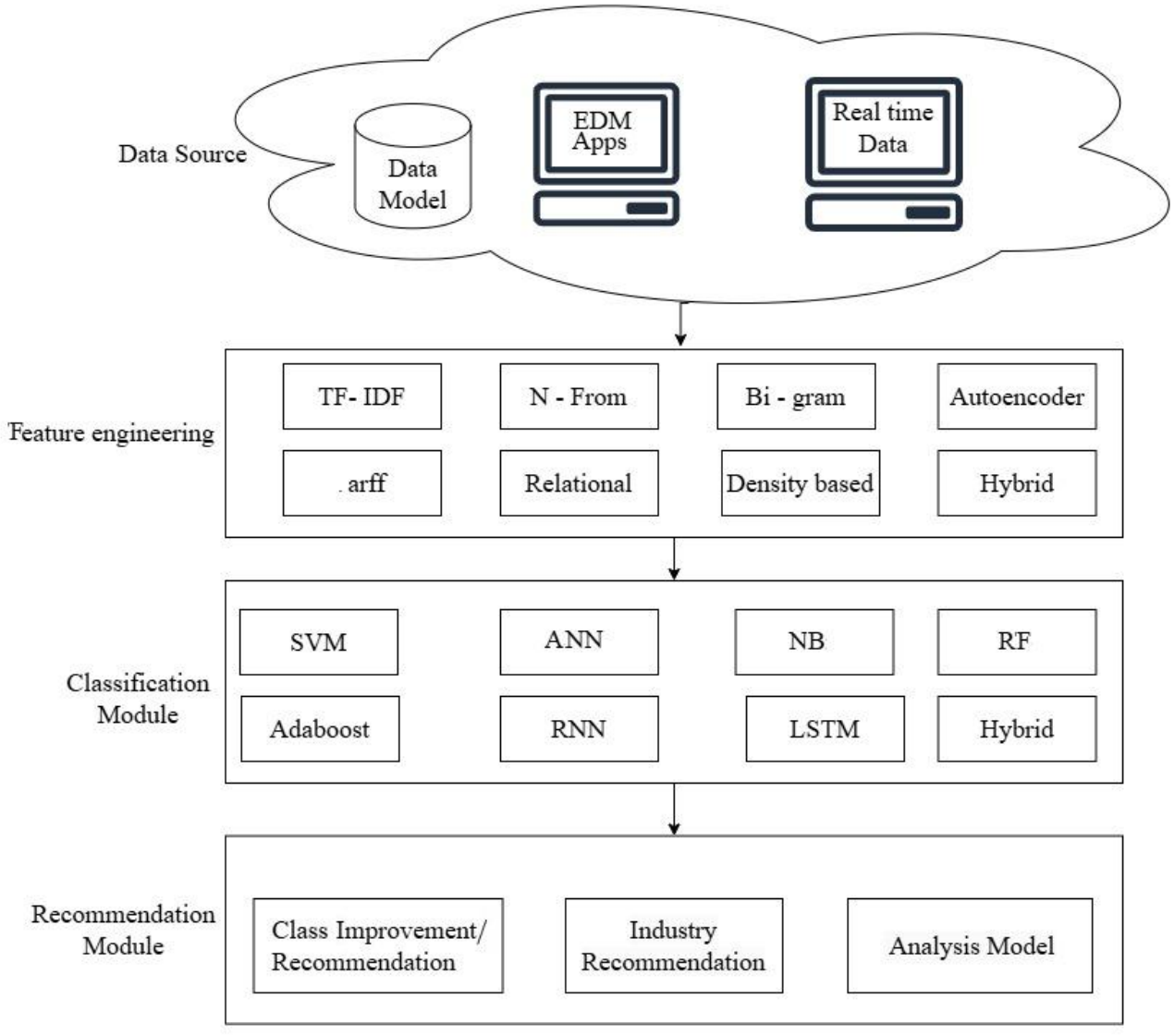

Figure 1 : proposed system architecture for EDM using hybrid ML and DL

Classification : Classification in the hybrid deep learning framework utilizes a range of machine learning and deep learning models to categorize and predict educational outcomes. Naive Bayes (NB), a probabilistic classifier, is employed for tasks such as predicting student dropout rates or categorizing student performance levels based on prior data. Its simplicity and efficiency make it suitable for initial exploratory tasks. Support Vector Machines (SVM), known for their ability to handle high-dimensional data, are often applied in identifying patterns such as learning difficulties or behavioral clusters among students. Artificial Neural Networks (ANN) take this further by capturing complex relationships in data, such as identifying latent factors influencing student performance. Random Forest (RF), an ensemble learning technique, is ideal for handling diverse data types. For example, it can classify students based on their performance in various subjects, leveraging feature importance measures to interpret results. Adaboost is often used in tandem to improve weak classifiers, boosting accuracy in predicting outcomes like test scores or course completions.

In deep learning, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks excel at modeling sequential data, such as tracking student progress over time. RNNs are particularly effective for time-series data, while LSTMs handle long-term dependencies, capturing sustained trends in learning. The hybrid approach, such as combining RNN with SVM, leverages the sequential modeling power of RNN with the classification strength of SVM. This approach enhances performance in tasks like predicting course success rates or identifying at-risk students, ensuring the strengths of each technique are fully utilized.

Recommendation Module : The recommendation module focuses on improving classroom practices and providing personalized industry recommendations. For classroom improvement, the system analyzes patterns in student engagement, feedback, and performance to suggest tailored interventions. For instance, it can recommend adaptive teaching strategies for specific groups or highlight areas where instructional content needs refinement. By leveraging hybrid deep learning, these recommendations are contextually aware, ensuring relevance to the unique needs of each classroom.

In addition, the module can identify effective peer learning opportunities, suggesting groups of students who can collaborate based on complementary strengths. It can also recommend targeted resources, such as additional practice material, video tutorials, or personalized assignments, to address individual learning gaps. On the industry side, the module focuses on career and skill recommendations, aligning educational outcomes with workforce demands. By analyzing student interests, performance data, and industry trends, it can suggest suitable career paths, certifications, or internships. For instance, a student excelling in programming might receive recommendations for software development roles, alongside courses on advanced coding techniques.

The hybrid deep learning approach ensures that recommendations are accurate, timely, and tailored to individual needs. By integrating insights from classification models, feature engineering, and real-time data, the recommendation module delivers actionable insights that benefit students, educators, and institutions. These personalized suggestions ultimately improve educational practices and help bridge the gap between academia and industry demands.

Algorithm Design

The hybrid deep learning framework for EDM and improving educational practices can be represented mathematically as a sequence of operations that combine feature engineering, dimensionality reduction, and hybrid classification models.

1. Feature Engineering and Data Representation

TF-IDF for Textual Data Representation

The Term Frequency-Inverse Document Frequency (TF-IDF) score for a term t in a document d within a corpus D is computed as:

$$TF\text{-}IDF(t,d)=TF(t,d)\cdot IDF(t,D)$$

where:

$$\mathrm{TF}(t,d)=\frac{f_{t,d}}{\sum_{t'\in d} f_{t',d}}\,(term\ frequency)$$

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{1+|\{d \in D: t \in d\}|}\right)_{(Inverse\ document\ frequency)}$$

-Grams and Bigram Features

For a sequence of tokens *{w1,w2,...,wn}*, bigrams are generated as:

$$\text{Bigram Features}=\{(wi, wi+1): i \in [1, n-1]\}.$$

Autoencoder for Dimensionality Reduction

Autoencoders minimize the reconstruction los *LAE:*

$$LAE = \| X - X^{\|2}2$$

where *X* is the input data, and *X^=f(g(X)* is the reconstructed data from the encoder *(g)* and decoder *(f)*.

Density-Based Clustering for Data Patterns

The SCAN algorithm identifies clusters based on a density threshold:

$$Core\ Distance(p)=\backslash minPts\ within\ \epsilon\text{-}neighborhood\ of\ p.$$

2. Classification Models

Recurrent Neural Network (RNN)

The RNN processes sequential data using hidden states:

$$h_t = f(W_h h_{t-1} + W_x x_t + b_h)$$

where:

- h_t is the hidden state at time t,
- W_h W_x are weight matrices,
- x_t is the input at time t,
- b_h is the bias term.

Support Vector Machine (SVM)

SVM optimizes the hyperplane for classification:

$$min \| w \| + C \sum_{i=1}^{n} n\xi i,$$

subject to:

$$y_i(\text{w}\cdot x_i + \text{b}) \geq 1 - \xi_i \text{i}, \xi_i \geq \text{o},$$

where:

- w is the weight vector,
- b is the bias,
- ξi are slack variables, and
- C controls the trade-off between margin width and classification error.

Hybrid RNN-SVM Model

The RNN-SVM hybrid combines RNN's sequential learning and SVM's classification as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b_h)\ , y{=}SVM(h_t)$$

where h_t is the final RNN hidden state passed to the SVM for classification.

Recommendation Module

The recommendation system generates actionable insights using weights WrW_rWr derived from the hybrid model:

$$R = w_r \cdot Y,$$

where:

- R represents recommendations,
- Y^ is the predicted outcome from the classification phase,
- w_r is a weighting matrix learned during training.

These algorithms collectively define the hybrid deep learning approach for Educational Data Mining. By leveraging these algorithms, the framework improves predictive accuracy and supports actionable insights to enhance education practices.

Results and Discussion

The confusion matrix analysis for the hybrid RNN-SVM classifier using the sigmoid activation function demonstrates its performance across three cross-validation scenarios: 5-fold, 10-fold, and 15-fold. The results highlight the model's consistent improvement in accuracy, precision, recall, and F1 score as the number of folds increases.

Table 1 : Confusion matrix analysis for RNN-SVM using sigmoid function

	5-Fold	10-Fold	15-Fold
Accuracy	0.9340	0.9534	0.9643
Precision	0.9666	0.9786	0.9810
Recall	0.9624	0.9725	0.9819
F1 Score	0.9645	0.9756	0.9815

In Figure 2 and Table 1 the model achieves 93.40% in 5-fold cross-validation, which further improves to 95.34% in 10-fold and 96.43% in 15-fold, showcasing the classifier's ability to correctly predict outcomes across varied datasets with minimal errors. Precision, a metric indicating the proportion of true positives among all predicted positives, starts at 96.66% for 5-fold and increases to 97.86% for 10-fold and 98.10% for 15-fold, highlighting the model's efficiency in minimizing false positives as data distribution becomes more refined.

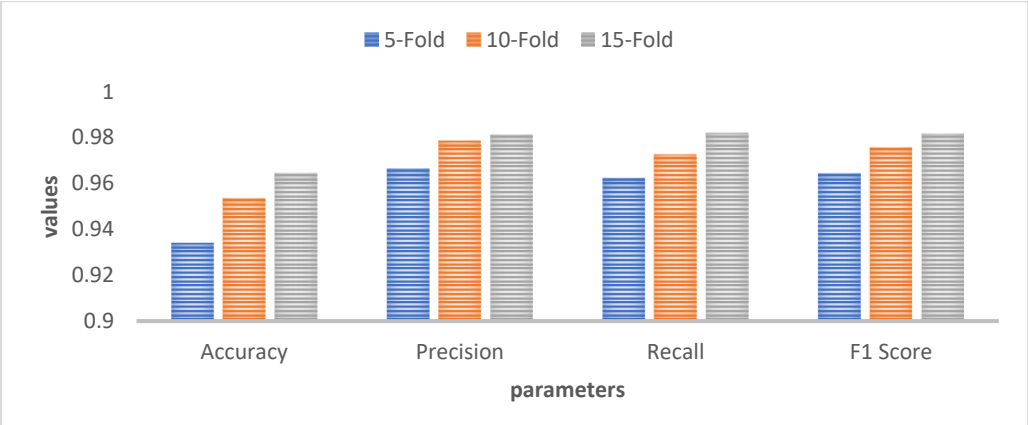

Figure 2 : Emotion classification accuracy of proposed model with hybrid feature extraction using sigmoid function

Similarly, recall, which measures the proportion of true positives identified out of all actual positives, also improves steadily, from 96.24% in 5-fold to 97.25% in 10-fold and 98.19% in 15-fold. This increase reflects the model's growing capability to capture all relevant instances in the data. Finally, the F1 score, a harmonic mean of precision and recall, follows a similar trend, starting at 96.45% in 5-fold, rising to 97.56% in 10-fold, and reaching 98.15% in 15-fold, indicating balanced and robust performance across these evaluation metrics.

Table 2: Confusion matrix analysis for RNN-SVM using Tanh function

	5-Fold	10-Fold	15-Fold
Accuracy	0.9526	0.9619	0.9660
Precision	0.9750	0.9795	0.9847
Recall	0.9750	0.9804	0.9800
F1 Score	0.9750	0.9800	0.9824

In Table 2 and Figure 3 achieves 95.26% in the 5-fold setup, which further improves to 96.19% and 96.60% for 10-fold and 15-fold validation, respectively. This upward trend highlights the model's scalability and capacity to generalize better with increased data splits. Precision, a critical metric for evaluating the correctness of positive predictions, also remains high, starting at 97.50% for 5-fold validation and climbing to 98.47% in the 15-fold configuration. This improvement signifies the model's ability to minimize false positives effectively.

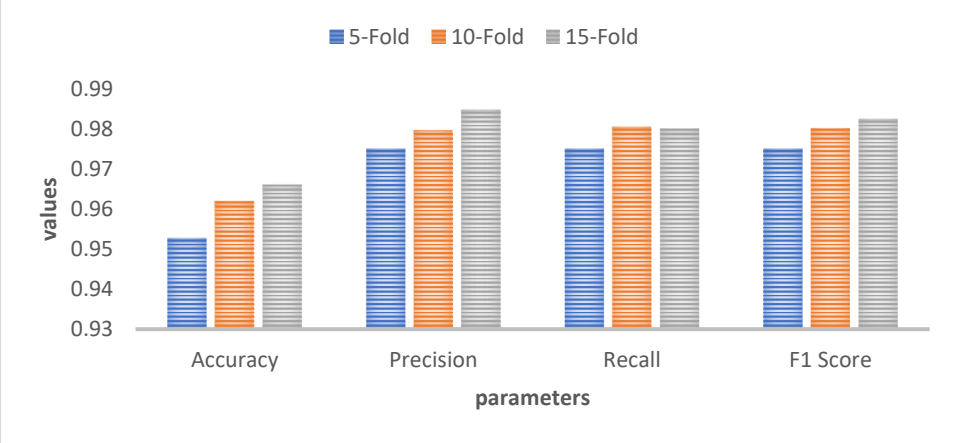

Figure 3 : Emotion classification accuracy of proposed model with hybrid feature extraction using Tanh function

Similarly, the recall metric, which measures the model's ability to identify true positives, maintains consistently high values across all folds. Starting at 97.50% for 5-fold validation, it reaches 98.00% for 15-fold validation, showcasing the model's aptitude in detecting relevant instances without significant drop-off. The F1 score, which balances precision and recall, follows a comparable trajectory, starting at 97.50% for 5-fold and peaking at 98.24% for 15-fold validation. This metric underscores the overall effectiveness of the hybrid RNN-SVM classifier in balancing false positives and false negatives.

Table 3 : Confusion matrix analysis for RNN-SVM using ReLu function

	5-Fold	10-Fold	15-Fold
Accuracy	0.9775	0.9748	0.9714
Precision	0.9926	0.9874	0.9921
Recall	0.9846	0.9867	0.9786
F1 Score	0.9885	0.9870	0.9853

In Table 3 and Figure 4 accuracy, which measures the overall correctness of predictions, the model achieves scores of 0.9775, 0.9748, and 0.9714 for 5-fold, 10-fold, and 15-fold cross-validation, respectively. These results highlight the model's ability to maintain high accuracy levels, even as the number of folds increases. Precision, which evaluates the proportion of true positives out of all predicted positives, remains exceptionally high across all folds. The scores are 0.9926 for 5-fold, 0.9874 for 10-fold, and 0.9921 for 15-fold validation, indicating the model's effectiveness in minimizing false positives.

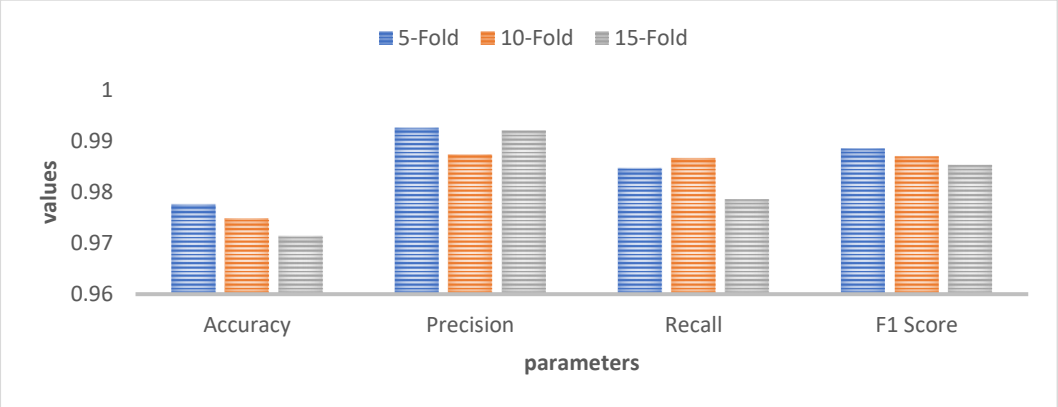

Figure 4: Emotion classification accuracy of proposed model with hybrid feature extraction using ReLu function

Recall, which measures the model's ability to identify true positives, is slightly lower than precision but still strong. The values are 0.9846, 0.9867, and 0.9786, suggesting the model's capacity to identify most relevant instances while maintaining consistency across folds. The F1 score, a harmonic mean of precision and recall, reflects the model's balanced performance. It records values of 0.9885, 0.9870, and 0.9853 for the respective folds, reinforcing the model's effectiveness in managing the trade-off between precision and recall.

Conclusion

The proposed EDM classification using hybrid learning techniques and advanced algorithms like RNN-SVM, combined with feature selection methods such as TF-IDF, Lemmatization, N-grams, and correlation-based techniques, has emerged as a highly effective approach for analyzing social media datasets. The RNN- SVM model excels at capturing sequential dependencies, while robust feature selection optimizes accuracy and minimizes computational overhead. TF-IDF emphasizes the most relevant terms in a document, while Lemmatization reduces redundancy by normalizing words to their root forms, improving generalization. N-grams enrich contextual understanding by capturing semantic relationships between words, and correlation-based feature selection filters out irrelevant features, enhancing model efficiency. The hybrid approach synergizes classical machine learning and deep learning methods, leveraging the model's ability to handle temporal dependencies for nuanced sentiment analysis. This combination consistently outperforms standalone methods in terms of accuracy, precision, and recall, particularly when processing noisy and high-dimensional social media data. Despite its strengths, challenges like overfitting and the computational demands of large datasets persist. Future research could address these challenges by optimizing training processes and exploring lightweight architectures for real-time applications. The study utilized 4000 records collected from Kaggle and real-time student feedback. Data preprocessing included tokenization, lemmatization, stopword removal, case normalization, and punctuation removal. Feature extraction involved Lemmas, N-grams, Bi-grams, NLP-based relationships, and hybrid lexicon features. Experiments were conducted to evaluate conventional classifiers, such as Naïve Bayes, SVM, and Artificial Neural Networks, alongside the proposed Hybrid Deep Learning (HML) approach. Sentiment analysis classified feedback into negative, positive, and neutral categories, with HML achieving 99.45% accuracy using hybrid feature selection. The RNN-SVM hybrid model demonstrated 97.75% accuracy. Future advancements could focus on lightweight, real-time models for large-scale datasets, integrating transformers and graph neural networks for context-aware embeddings, and enhancing cross-linguistic and cross-domain generalization for broader applicability.

References

[1] Usman Ali, Khawaja Sarmad Arif and Dr. Usman Qamar. "A Hybrid Scheme for Feature Selection of High Dimensional Educational Data", 2019, International Conference on Communication Technologies (ComTech 2019), IEEE.

[2] Gabriella Casalino, Giovanna Castellano, Andrea Mannavola and Gennaro Vessio. "Educational Stream Data Analysis: A Case Study", 2020, IEEE.

[3] Bhanuka Mahanama, Wishmitha Mendis, Adeesha Jayasooriya, Viran Malaka, Uthayasanker Thayasivam and Thayasivam Umashanger. "Educational Data Mining: A Review on Data Collection Process", 2018, International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE.

[4] Michal Munk, Martin Drlík, Ľubomír Benko and Jaroslav Reichel. "Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques", 2017, IEEE.

[5] Akansha Mishra, Rashi Bansal and Dr. Shailendra Narayan Singh. “Educational Data Mining and Learning Analysis”, 2017, IEEE.
[6] John Jacob, Kavya Jha, Paarth Kotak and Shubha Puthran. “Educational Data Mining Techniques and their Applications”, 2015, IEEE.
[7] El Harrak Othman, Slimani Abdelali and El Bouhdidi Jaber. “Education Data Mining: Mining Moocs Video using Metadata Based Approach”, 2016, IEEE.
[8] Lixia Ji, Xiao Zhang and Lei Zhang. “Research on the Algorithm of Education Data Mining Based on Big Data”, 2020, 2nd International Conference on Computer Science and Educational Informatization (CSEI), IEEE.
[9] Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan. “A Systematic Review on Educational Data Mining”, 2016, IEEE.
[10] J. L. C. Ramos, R. E. D. Silva, R. L. Rodrigues, J. C. S. Silva and A. S. Gomes. “A Comparative Study between Clustering Methods in Educational Data Mining”, 2016, IEEE.
[11] Manas Chaturvedi. “Data Mining and it’s Application in EDM Domain”, 2017, International Conference on Intelligent Computing and Control Systems ICICCS, IEEE.
[12] Camilo Vieira, Alejandra J. Magana and Mireille Boutin. “Using Computational Methods to Analyze Educational Data”, 2019, IEEE.