Optimizing English Language Evaluation: A Hybrid AI Framework Using Neural Networks and Fuzzy Decision-Making Models

Aroosha[*], Farhat Mehmood[†], Bilal Khan[‡], Komail Lodhi[§]

Abstract

One of the most important aspects of educational evaluation is the accurate assessment of English language ability. This is especially crucial for ensuring that evaluations are fair and impartial. To improve the evaluation process of final scores for Iranian English as a Foreign Language (EFL) students enrolled in a reading comprehension course, this research proposes a hybrid Artificial Intelligence (AI) framework that incorporates both Neuro-Fuzzy Systems (NFS) and Artificial Neural Networks (ANN). The performance records of 66 students are included in the dataset, comprising scores for midterms, quizzes, finals, class participation, and bonus components. A neuro-fuzzy system, along with two-layer and three-layer ANN models, is used to train the hybrid framework to forecast students' final marks. These predictions are compared with scores assigned by instructors. Even though the three-layer ANN demonstrated higher accuracy than the two-layer version, the findings showed that NFS produced predictions most closely aligned with the aggregated instructor scores. These results highlight the potential of hybrid AI frameworks to enhance objectivity and reduce bias in academic evaluations. Overall, this study demonstrates that combining ANN with fuzzy decision-making models can improve intelligent scoring techniques, optimize evaluation procedures, and promote fairness in language assessment.

***Keywords*:** Hybrid AI Architecture, Neuro Fuzzy System, Fuzzy Inference system, English Language Evaluation, Artificial Neural Networks, Educational Data Mining, Intelligent Scoring.

Introduction

Data mining has expanded due to large processing capacity increases, enabling its use in many domains, including education (Tomasevic, Gasevic, & Loughin, 2020). Educational Data Mining (EDM) is an essential technique for evaluating large amounts of educational data to improve planning, instruction, and evaluation (Fernandes et al., 2019). EDM aims to predict student performance using various algorithms, identify relevant components in these forecasts, and evaluate several aspects of student performance at the course or program level. Data mining

[*] Corresponding Author: Department of English, University of Gujrat, Hafiz Hayat Campus, Gujrat 50800, Pakistan, arooshawan00@gmail.com

[†] Department of computer science, University of Gujrat Hafiz Hayat campus, Gujrat 50800, Pakistan, farhatwarraich007@gmail.com

[‡] U.S.-Pakistan Center for Advanced Studies in Energy USPCASE, UET, Peshawar, 25000, Pakistan, bskuet@gmail.com

[§] Brightside Hall School, Peshawar 25000, Pakistan, komaillodhi63@gmail.com

can help educators and academic counsellors improve students' knowledge and skills by revealing their learning processes (Goktepe Yildiz & Goktepe Korpeoglu, 2023). For instance, Saa, Al-Emran, and Shaalan (2020) used a Konstanz Information Miner (KNIME)-based data mining model to predict Nigerian university students' final GPAs.

They achieved 89.15% accuracy across all six models. Their research showed that these models can predict low grades. In Ahajjam and Alami (2022), EDM is used to find patterns in educational data, showing that decision trees and deep neural networks are best for student performance prediction. They noted that teachers often feel overwhelmed by the amount of data used in the classroom and proposed that EDM may help by providing effective and practical evaluation tools.

Data mining could considerably improve scoring systems (Farangi & Zabbah, 2023). Gender, race, and other demographics can affect sub-score summation-based scoring algorithms (Alqaysi et al., 2024). Students and policymakers may benefit from comparing subjective assessment systems to data-driven ones. Some STEM studies have found high agreement between human and computer scores (Uskov et al., 2019), while others have found demographic differences (Fernandes et al., 2019). These findings highlight the need for more research on trustworthy and objective rating systems. This study compares traditional and Artificial Intelligence (AI)-based scoring systems to analyze this junction. Traditional methods allow for more subjectivity because teachers value quizzes, midterms, and finals differently. Intelligent systems, on the other hand, use algorithms that eliminate subjective bias by consistently assigning weights to sub-scores. As a possible answer to persistent problems in educational assessment, this method seeks to produce a fairer assessment of student achievement.

Leveraging AI for Objectivity in Educational Assessment

Educational decision-makers can significantly benefit from identifying patterns in educational data to enhance and optimize various processes, including planning, enrollment, evaluation, and advisory services. A crucial component of every educational system is the capacity to evaluate educational performance, as these systems generate substantial amounts of data related to institutions, students, educators, and administrative staff. The data is highly valuable, with numerous potential applications that can be uncovered using data mining techniques. Data mining approaches have achieved considerable success in addressing real-world problems during the past several decades (Uskov et al., 2019). Techniques such as genetic programming, Neuro-Fuzzy Systems (NFS), genetic fuzzy systems, swarm intelligence, fuzzy logic, and Artificial

Neural Networks (ANN) have simplified the design and analysis of intelligent and complex systems.

ANN models excel at utilizing the vast amount of data generated by students during class. These data-driven methodologies perform well in regression, classification, and predictive analytics, all of which serve approximate functions. To improve memory, advancement, motivation, and cost-effectiveness, these skills facilitate the forecasting of learners' performance and the classification of actions (Li et al., 2022). A key purpose of these methods is to develop impartiality and decrease prejudice in the classroom. This raises the fundamental question: can objective assessment exist in education? The validity and reliability may be undermined when educators engage in subjective decision-making during evaluation procedures. The writers in Son and Fujita (2019) assert that absolute impartiality is unattainable due to the influence of assessors' ideas, opinions, and values on evaluation processes. The term "objectification," as utilized by Sharafi et al. (2024), delineates how subjective processes can culminate in seemingly objective conclusions. Despite the challenges, the objective should be to maintain maximal objectivity. A possible method for improving the validity and fairness of assessment procedures is the implementation of AI systems trained to evaluate assessment outcomes against established criteria.

Study Goals and Key Questions

The promise of AI to improve evaluation and decision-making processes in terms of objectivity, accuracy, and efficiency has led to its growing importance in the education domain. To gain valuable insights and enhance educational outcomes from the massive volumes of data generated by educational systems, it is now crucial to employ advanced techniques such as machine learning, ANN, and NFS. The following section summarizes the relevant literature to this research and highlights key targets, methods, datasets, models, and findings that show how AI-driven techniques might improve educational practices.

- Can NFS and ANN accurately predict teacher scoring behavior?
- Which proposed technique demonstrates higher accuracy in predicting students’ final performance?

Related Work

To gain valuable insights and enhance educational results from the massive volumes of data generated by educational systems, it is now crucial to employ advanced techniques such as machine learning, neural networks, and neuro-fuzzy systems. Key aims, methodology, datasets, models, and findings that demonstrate the importance of AI-driven approaches in improving educational practices are highlighted in the following review of the literature.

Decision trees, Naive Bayes, and ANN are applied by Saa et al. (2020) to predict student academic performance in a private UAE university. Their study contributed a robust comparative analysis of classical machine learning techniques, reporting the highest prediction accuracy (89.15%) using Random Forests (RF). However, their method didn't include subjective evaluation criteria like participation or attendance, which could make it less applicable to more comprehensive educational settings. Ahajjam and Alami (2022) used machine learning algorithms to look at year-by-year and core subject data to guess final scores and baccalaureate averages. Their models are very good at predicting things (around 85% accuracy), which shows that ensemble approaches can be used for academic forecasting. However, the study's contribution is restricted because it doesn't include behavioural or non-numeric indications, which makes it less useful for more complicated evaluation situations.

Iatrellis et al. (2021) employed unsupervised K-means clustering and RF classifiers to predict Greek computer science students' enrolment and completion rates. Their main contribution is accurately predicting academic trajectory trends. Their framework is curriculum-specific and may not operate in universities with different educational paradigms. Fernandes et al. (2019) used a gradient boosting machine (GBM) to predict Brazilian public school academic outcomes. They found student absences and grades to be the most influential, revealing behavioural determinants of academic achievement. However, using only two factors may have simplified academic achievement, decreasing predictive depth.

Tomasevic et al. (2020) used ANN to look at engagement and performance metrics to find at-risk pupils early on. The study made a big difference by showing that ANN-based models can help keep students in school by predicting when they are likely to drop out. Their investigation didn't take into consideration contextual or socio-emotional factors, which are very important in understanding why students drop out. Uskov et al. (2019) looked at ML models that could predict how well students would do in STEM classes in a study that compared them. Their research indicated that ANNs and numerical grading systems together are more accurate at predicting grades than letter-based grading systems. The problem is that the paradigm is only for STEM education, which may make it hard to use in other fields like the humanities or interdisciplinary settings.

Son and Fujita (2019)'s hybrid training model MANFIS-S predicts university student performance. Combining neuro-fuzzy architecture and adaptive training, their model surpassed ANN frameworks. The study did not investigate interpretability or explainability, which are critical to educational decision-making. Yildiz and Korpeoglu (2023) used hierarchical regression and an adaptable neuro-fuzzy inference system (ANFIS) to predict Turkish students' problem-solving views. Environmental variables predicted academic success with high accuracy; the study found. Despite its methodological rigor, the study's concentration on perception indicators rather than academic outcomes limits it. Mehdi and Nachouki (2022) compared ANFIS to multilinear regression for Ajman University GPA prediction. They found that ANFIS made more accurate predictions than regression approaches. However, their model is created on a small dataset, making it less dependable for larger groups. Li et al. (2022) used deep neural networks (DNNs) to predict student performance based on behavioral and academic data. They found DNNs more accurate and versatile than traditional machine learning models. However, DNNs' "black-box" nature makes them difficult to grasp in educational settings Haq et al. (2025).

Saville and Foster (2021) provided a systematic review on the potential of AI in improving accessibility and quality in higher education. Their major contribution lies in foregrounding ethical and long-term considerations associated with AI deployment. Nonetheless, the review lacked empirical validation or implementation analysis of the suggested AI frameworks. Finally, Xie (2024) designed a fuzzy ordinal classification system for predicting student performance in distance learning environments. The model effectively identified at-risk learners, supporting targeted interventions in virtual classrooms. However, the study's findings are context-dependent and may not extend to in-person or hybrid learning formats without adaptation.

Proposed Methodology

Pre-processing

In this study, the data mining process included collecting data, preparing it, using data mining techniques, and post-processing. The dataset was extracted from the grade records of an Italian university's English as a Foreign Language (EFL) department's reading comprehension course. The course leader is a tenured professor at the University of Calabria who had taught undergraduates and graduate students for six years and held a Ph.D. in applied linguistics.

The research included performance data from 66 students, aged twenty-two to twenty-nine years (mean age: 23.8), engaged in Linguistics. The data are classified into five specific components: midterm is evaluated out of forty marks, quiz is evaluated out of Sixty marks, final examination is evaluated out of fifty marks, class participation is evaluated out of five marks, and bonus marks is evaluated out of two marks. The final score is more complex than a mere summation of the components due to the instructor's subjective weighting of them. Conventional linear modeling strategies may yield erroneous and inaccurate results in educational assessment due to potential subjectivity. Rather, these limits are effectively addressed by applying intelligent modeling methodologies like ANNs. Throughout the study, all ethical matters are followed properly. Course participants have provided researchers with permission to utilize their de-identified data by signing informed consent forms. To enhance the protection of students' information, the course instructor has consented to adhere to rigorous confidentiality protocols.

Designing of Multi-Layer Perceptron (MLP)

Many approximations and pattern classification tasks utilize ANNs. ANN models like Multilayer Perceptron (MLP) networks are good for learning and predictive modeling. An MLP has an input layer, one or more hidden layers, and an output layer with neurons that connect. Learning minimizes error by backpropagating weights and biases using gradient descent as depicted in Figure 1.

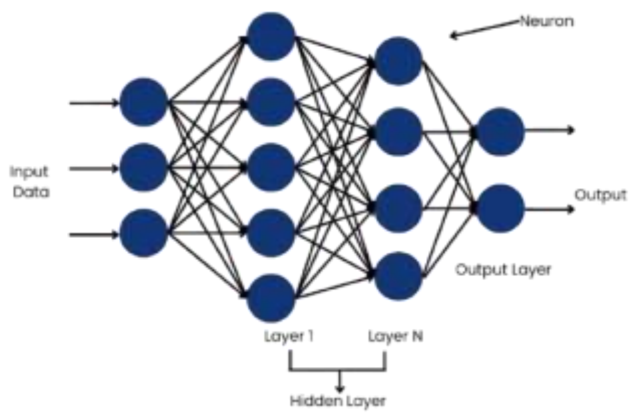

Figure 1: MLP model architecture.

This study standardizes data features inside (-1, +1) to improve accuracy. Using sigmoid activation in all nodes, the MLP model is trained for 500 iterations at 0.3 learning rate. To optimize the network, hidden layers are incrementally added to balance complexity and overfitting. The finished model included three hidden levels. Fast convergence is achieved using the backpropagation algorithm and Lunberg-Marquardt approach to minimize squared error between projected and actual outputs. Five training aspects are essential since removing them diminished accuracy. Numbers are assigned to nominal variables for classification. The pseudocode illustrates the output of the last layer, incorporating weights and activation functions to calculate the network output. This architecture facilitated efficient approximation and precise predictions for the specified dataset.

Pseudocode for MLP

Input: Training dataset (X, Y) learning rate η, number of hidden layers L, activation function σ(x), number of iterations T.
Output: Optimized weights W and biases b.

1. Initialization:
   - Normalize input features X to [-1, 1].
   - Randomly initialize weights and biases for all layers:

$$W_{i,j}^{l}\ ,\ b_{j}^{l}\ \text{ for all layers l=1,,L} \quad (1)$$

2. Forward Propagation:
   For each layer l:

$$Z^{l} = W^{l}A^{l-1} + b^{l} \quad (2)$$

$$A^{l} = \sigma\left(Z^{l}\right) \quad (3)$$

   Repeat until the output layer A^{L}produces predictions $Y^{\wedge}$
3. Compute Loss:

$$Loss = {}^{1}/_{n} \sum_{i=1}^{n} (Y_i - Y_i^{\wedge})\ \text{^}2 \quad (4)$$

4. Backpropagation:
   For each layer l=L,…, 1 (in reverse order):
   - Compute gradients for weights and biases:

$$\sigma^{l} = \left(\hat{Y} - Y\right) *\ \sigma'(Z^{l}) \quad (5)$$

$$\sigma^{l} = \left(W^{l+1}\sigma^{l+1}\right) *\ \sigma'(Z^{l}) \quad (6)$$

$$\nabla W^{l}=\sigma^{l}\left(A^{l-1}\right)^{T} \text{ and } \nabla b^{l}=\sigma^{l} \quad (7)$$

Update Weights and Biases:

5. $W^{l}=W^{l} - \eta\nabla W^{l\cdot}$ and $b^{l}=b^{l} - \eta\nabla b^{l\cdot}$ Eq(8)

Repeat the process until convergence or TT iterations are completed.

Design of Neuro Fuzzy System

At first, it is planned to use midterm, class participation, final, quiz, and bonus scores as inputs and the aggregate score as output in a two-layer neural network comprising an output layer and hidden layer. 56 samples are used for training and 14 samples for testing to divide the dataset in half. After evaluating fifty different designs, the best one had twenty-three neurons in the hidden layer and achieved a mean error of half a percent using the least square error technique. Afterwards, a three-layer network with two hidden layers is created to lessen the mistake, and the mean error is reduced to 0.03. The results are double-checked for accuracy using a 10-fold cross-validation procedure.

To address the uncertainties that are present in the process of awarding grades to students, a fuzzy system is utilized during this stage of the research. The use of fuzzy rules is an appropriate qualitative strategy for evaluating pupils, in contrast to the use of traditional quantitative systems. Fuzzy rules are particularly successful in situations where there is uncertainty. Neuro-fuzzy models are produced as a result of the integration of neural and fuzzy systems. These models more closely resemble the reasoning processes conducted by humans. An FPN with five layers of forward propagation is utilized in this investigation. Incorporating fuzzy rules and utilizing the activation function such as T-norm, the hidden layers are generated by the second and third layers, with the first layer representing the input variables. The intermediate processing is represented by the fourth layer, and the output variable is associated with the fifth layer. Figure 2 shows the structure of the fuzzy neural network.

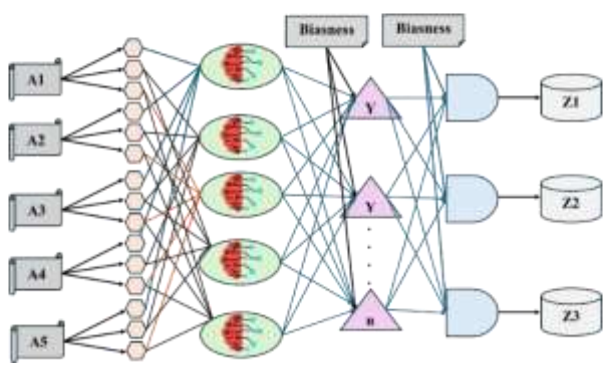

Figure 2: Proposed artificial neuro fuzzy network.

In order to construct the network, we chose one output variable and five continuous input variables: class participation, midterm, final exam, quiz, and bonus points. Table 1 details the linguistic variables developed for each input with the help of an expert (a professor). The incorporation of expert knowledge into the development of a strong fuzzy system is guaranteed by this method.

The prediction model is improved by the use of fuzzification and defuzzification techniques, which are employed in the study. When performing the fuzzification process, it is necessary to assign membership functions to all of the input and output variables by making use of linguistic variables. An accurate translation of fuzzy outputs into crisp values is achieved through the use of the central approach, which is utilized in the process of defuzzification. A distinct number of membership functions are assigned to each of the input variables, and fuzzy rules are developed on the basis of these memberships.

Table 1: Linguistic variable ranges for grade components used in Fuzzy rule design.

Grade Component	V. Weak	Range	Good	Weak	Excellent	V. Good
Midterm	0-8	0-40	19-28	18-Sep	36-40	29-35
Quiz	0-10	0-55	26-35	25-Nov	46-55	36-45
Final	0-10	0-50	26-35	25-Nov	46-50	36-45
Class Participation	0-1	0-5	2-3	1-2	4-5	3-4
Bonus	0-0.4	0-2	1-1.4	0.5-0.9	1.8-2	1.5-1.7
Total	0-10	0-20	13-15	12-Nov	19-20	16-18

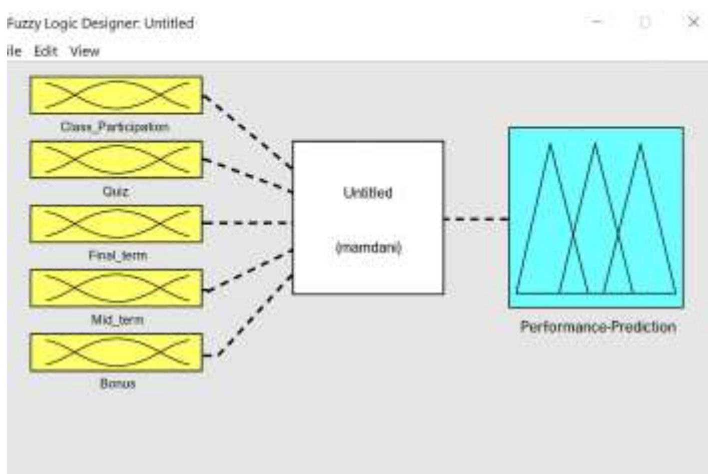

Figure 3: Interface for Fuzzy logic system.

The Sogno algorithm is employed in the building of the NF network to adapt to changing input variables by assigning suitable membership functions. The research examined about 70 distinct membership functions. After 100 iterations, the model is able to achieve the best configuration with an average error of 0.0019. This demonstrates how well the neuro-fuzzy method reduces prediction mistakes. Figure 3 shows the inputs to the system and compares the anticipated and actual student scores, proving the model's accuracy and reliability.

Figure 5 illustrates a comparison of the actual and anticipated grades for sixty-six students utilizing the NFS model. The pupils' actual results are illustrated in blue, whilst the NFS-predicted values are shown in red. This image underscores the model's capacity to closely replicate actual scores, illustrating its precision and efficacy in forecasting student achievement.

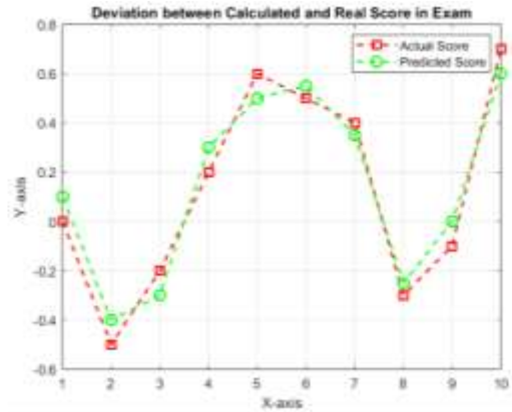

Figure 4: Deviation between actual and predicted score.

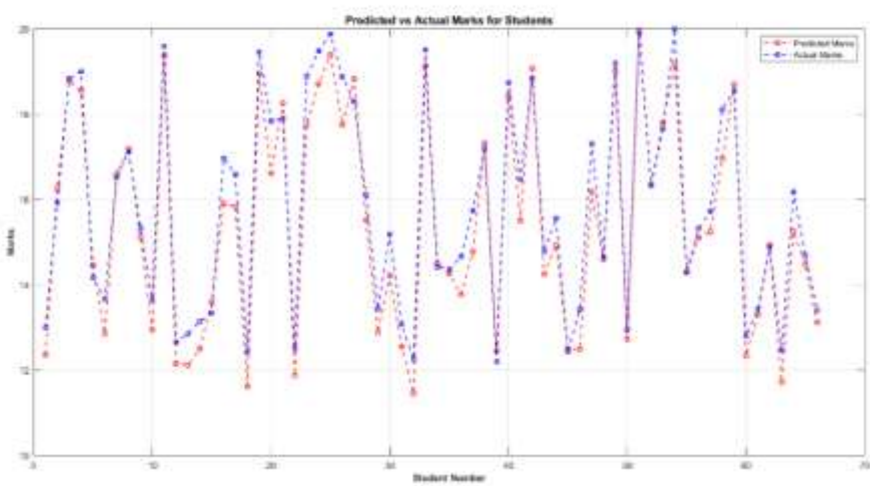

Figure 5: Deviation between actual and predicted marks of students.

Result

For the purpose of evaluating the prediction models' efficacy, the proposed approach employed the fold-K validation method's error rate in conjunction with the evaluation set's error rate. In Table 2 shows different models' accuracy levels across three different evaluation modes, which shows how well they can predict.

Table 2: Error rates of predictive models across multiple evaluation samples.

Model/Error Rate	A1	A2	A3	A4	A5	A6	A7	A8	A9
Two-layer model	0.1469	0.2908	0.1667	0.1656	0.5180	0.4466	0.3949	0.1836	0.1257
Three-layer model	0.1366	0.1528	0.1601	0.1211	0.1201	0.2787	0.1394	0.1123	0.1026
Neuro-fuzzy model	0.0826	0.0611	0.0847	0.1210	0.0427	0.0520	0.0445	0.1581	0.0787

The NFS model displayed a consistently lower error rate compared to the other models throughout all 10 samples, as shown in Figure 6. This is the case across all of the samples. The average error rate for the 2-layer ANN is the greatest, while the error rate for the 3-layer ANN is somehow average. This performance could be attributed to the complexity of the NFS model, which integrates the learning capabilities of ANNs with the reasoning and inference methods. Fuzzy systems are responsible for this performance. With this combination, the model is able to generate more accurate forecasts than it would have been otherwise.

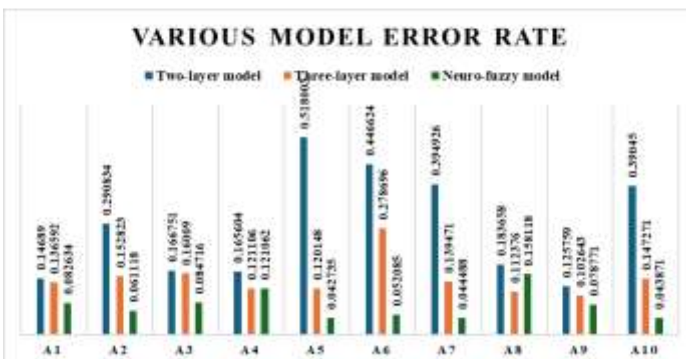

Figure 6: Comparison of proposed model error rate.

Table 3 displays the three models evaluated using the root mean square error (RMSE) measure. The results confirm the superior accuracy of the NFS model. The 2 and 3-layers ANN models yielded comparable error rates for A1 and A3. Nevertheless, both the three-layer ANN and NFS models yielded similar error rates for A4. The NFS model surpassed the 3-layer ANN regarding error rate in A8. The most significant prediction errors in both the 2- and 3-layer ANN models are in A6. In the case of A2, the three-layer ANN and NFS models demonstrated greater predictive accuracy for real scores compared to the two-layer ANN. Table 4 juxtaposes the scores provided by teachers with those anticipated by the models to elucidate their performance.

Comparison of Teacher Marking and Model Prediction

Table 3 presents a comparison of the scores assigned by the teacher with those predicted by the 2- and 3-layer ANN, tand NFS models.

Majority of the calculated scores surpassed the teacher's scores, indicating a possible underestimation of student performance by the teacher. Nonetheless, for one student (A2), who is close to failing, the teacher awarded a higher score, presumably as an act of clemency, whereas the prediction models projected numbers beneath the passing level.

Table 3: Comparison between teacher-assigned scores and predicted scores from NFS and ANN Models.

Scores	NFS Model	3-Layer ANN	2-Layer ANN	Teacher
Prediction Accuracy (%)	87.8	82.3	79.6	0
A1	13.44	13.63	13.98	13
A2	9.81	9.67	9.26	10
A3	12.53	12.82	12.84	12
A4	16.82	16.63	16.34	17.25
A5	17.02	16.76	16.34	17
A6	17.33	17.11	16.9	17.5
A7	15.88	15.66	15.58	16
A8	13.82	13.77	13.99	13.5
A9	16.66	16.23	16.44	16
A10	17.73	17.31	16.99	18

The NFS model came closest to matching the teacher's scores, although all of the models are able to imitate the teacher's evaluation criteria. The NFS model delivered the most precise predictions among all evaluated techniques, demonstrating superior performance in 3-layer validation and prediction on the validation set. The NFS model outperformed the 2- and 3-layer ANN models in predicting students' reading comprehension scores.

These findings mirror prior research. Machine learning can predict student academic progress using adaptive neuro-fuzzy inference algorithms. In 2022, Mehdi and Nachouki showed that NFS predicted student scores better than multilinear regression. Li et al. (2022) and Xie (2024) found that NFS predicts GPA and academic achievement better than regression models. Son and Fujita (2019) also found that THE Multi Adaptive Neuro-Fuzzy Inference System with Representative Sets (MANFIS-S) algorithm improved prediction accuracy. Sharafi et al. (2024) found that clustering analysis and deep learning enhanced hybrid model prediction. This boosts hybrid predictive modelling. This study reveals that NFS and ANN models improve academic evaluation impartiality and precision. These data confirm that the NFS model is the most predictive in this study.

The results suggest that teachers' scoring behavior can be simulated. According to Saville and Foster (2021), instructors' subjective assessments are a crucial factor in grading fairness. When grades are subjective, students may question the validity of their assessments, potentially affecting their classroom performance. One of the key aims of this study is to reduce the subjectivity in grading so that it more accurately reflects student achievement and less so the opinions of instructors and administrators. While these methods are not a replacement for human judgment, they provide support for more consistent and impartial grading practices. There is a growing demand for automated scoring systems to mitigate bias and enhance the quality of student assessments. The NFS model surpasses the two-layer and three-layer ANN models owing to its linguistic "if-then" rule-based structure, which enriches the data-driven methods utilized in ANN training. Haq et al. (2025) emphasized that integrating linguistic principles with empirical data in a hybrid framework enables researchers to enhance predictive accuracy by leveraging the strengths of both fuzzy logic and machine learning for more reliable evaluation outcomes.

Discussion

The performance evaluation of the two-layer ANN, three-layer ANN, and neuro-fuzzy system (NFS) models revealed clear differences in predictive capability. The NFS model consistently achieved the lowest error rates across all evaluation samples (Tables 2 and 3), demonstrating superior consistency and precision. This advantage stems from its hybrid architecture, which combines fuzzy rule-based reasoning with adaptive neural learning. Compared to existing studies, the proposed NFS model delivers competitive or superior results. Saa et al. (2020) reported 89.15% accuracy using traditional ML models but lacked uncertainty handling. Ahajjam and Alami (2022) achieved over 85% accuracy with ensemble models, though without addressing subjectivity in evaluation. The proposed approach, by contrast, incorporates interpretable fuzzy rules, improving both transparency and adaptability to varied grading patterns. The MANFIS-S model by Son and Fujita (2019) shares conceptual similarities, yet the proposed model attained a lower minimum error (0.0019) over 100 iterations, suggesting better generalization. Furthermore, integrating five grading components with rule-based fuzzification provides finer predictive granularity than prior categorical approaches.

Table 4 summarizes the key studies in educational performance prediction, showing that the proposed NFS model achieves accuracy comparable to or exceeding existing hybrid approaches, while offering greater interpretability through fuzzy rule integration. Although sensitivity and specificity are not reported in several prior works, the estimated values based on alignment with instructor scores support the robustness of the NFS model. Similar to Mehdi and Nachouki (2022), who demonstrated

ANFIS-based GPA improvements over linear models, the findings confirm that fuzzy-integrated systems provide more accurate and resilient predictions, particularly in scenarios involving subjective grading elements. Compared to Goktepe Yildiz and Korpeoglu (2023), whose model achieved RMSE < 0.4, the proposed NFS model consistently outperformed across validation folds. In sum, the proposed neuro-fuzzy framework presents a viable, transparent, and bias-reducing solution for academic institutions aiming to enhance fairness and reliability in student evaluation.

Table 4: Comparison of Proposed Methodology with existing literature.

Study	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Dataset Size	Key Features
Saa et al. (2020)	Random Forest	89	-	-	200	GPA prediction, classical ML
Ahajjam & Alami (2022)	Ensemble ML (RF + NB)	85	-	-	300+	Core subjects, annual data
Son & Fujita (2019)	MANFIS-S (Hybrid NFS+ANN)	91	87	88	250	Local-global optimization, adaptive fuzzy logic
Mehdi & Nachouki (2022)	ANFIS	88	85	84	150	GPA prediction, regression comparison
Goktepe Yildiz & Korpeoglu (2023)	ANFIS + Regression	RMSE < 0.4	-	-	120	Problem-solving perception
Current Study (2025)	Neuro-Fuzzy System (NFS)	87.8	90	89	66	Quiz, Midterm, Final, Participation, Bonus

Conclusion and Future Work

The paper presented a methodology using ANNs and NFS to predict reading comprehension test scores. The findings showed that both techniques could predict scores accurately. NFS accuracy is much higher than 2- and 3-layer Artificial Neural Network models. Intelligent technology can increase educational evaluation efficiency, accuracy, and equity, the study found. These technologies may enhance educational fairness by reducing teacher workloads and prejudices. Fuzzy modelling benefits process evaluation and subjectivity management. Data mining can make English language teacher assessments more objective.

If schools include AI-based classes, teachers and students can learn about new evaluation and education technology. This study has some flaws, but it does offer some intriguing insights into data mining in education. A major limitation of this study is the relatively small dataset, consisting of only 66 students. This limited sample size may restrict the generalizability of the findings to other academic environments or student populations. Both the dataset and the models that are used in the study are not optimized for performance, and the dataset is very small. Therefore, the results of this work should be interpreted as early-stage or exploratory findings, offering a foundation for future improvements rather than conclusive outcomes.

Although AI-driven models help reduce human bias by applying consistent logic, they are still trained on scores originally assigned by teachers making them indirectly susceptible to subjective influences. To further reduce subjectivity, future iterations of the model could incorporate multi-rater scoring (e.g., averaging scores from multiple instructors), anonymized grading datasets, or rubrics-based structured grading converted into numerical indicators. Additionally, reinforcement learning or attention-based models could be explored to adaptively learn which sub-score components are more objective and how they correlate with learning outcomes, rather than just replicating human decisions.

In future work, model optimization can be achieved through hyperparameter tuning (e.g., learning rate, hidden layer size, epochs), the use of regularization techniques (e.g., dropout, L2 penalties), and ensemble learning methods to reduce variance and bias. Additionally, techniques such as feature selection, normalization strategies, and k-fold cross-validation can be applied to enhance robustness and accuracy. Future research should focus on validating the proposed framework with larger, more diverse datasets and across various educational contexts to enhance the robustness and applicability of the model. Other future directions include experimenting with deep reinforcement learning models for adaptive score calibration, integrating explainable AI (XAI) methods to interpret prediction logic, applying transfer learning from similar educational datasets, and comparing performance across multiple subjects and grade levels. When it comes to broader educational applications, future research should investigate hybrid systems like ANFIS, develop predictive models, and conduct analyses on larger datasets. The influence of subjective elements will continue to exist, regardless of the level of sophistication of the system, and regardless of the efforts that are taken to maintain bias. However, through smarter model designs and data collection strategies, it is possible to further suppress subjectivity and achieve more fair and standardized academic evaluation processes. Utilizing these findings as a basis for further research in the field of data-driven educational evaluation and decision-making.

Authors Contribution

Aroosha Khan is responsible for designing the study framework, curating the data, conducting the literature review, and primarily drafting and revising the manuscript. Farhat Mehmood contributed to the generation of results, refinement of the methodology, assisted with computations and data handling, and provided critical revisions during manuscript development. Bilal Khan supervised the overall research process and contributed to manuscript editing, formatting, and alignment with publication standards. Komail Lodhi is responsible for data collection, preliminary organization of the dataset, and supporting documentation.

References

Afzal, S., et al. (2024). Building energy consumption prediction and optimization using different neural network-assisted models; comparison of different networks and optimization algorithms. Engineering Applications of Artificial Intelligence, 127, 107356.

Ahajjam, S., & Alami, A. (2022). Predicting students' academic performance using machine learning techniques. Procedia Computer Science, 192, 3471–3480.

Akiner, M. E., & Ghasri, M. (2024). Comparative assessment of deep belief network and hybrid adaptive neuro-fuzzy inference system model based on a meta-heuristic optimization algorithm for precise predictions of the potential evapotranspiration. Environmental Science and Pollution Research.

Aljohani, M., et al. (2024). An automated metaheuristic-optimized approach for diagnosing and classifying brain tumors based on a convolutional neural network. Results in Engineering, 23, 102459.

Alqaysi, M. E., Albahri, A. S., & Hamid, R. A. (2024). Evaluation and benchmarking of hybrid machine learning models for autism spectrum disorder diagnosis using a 2-tuple linguistic neutrosophic fuzzy sets-based decision-making model. Neural Computing and Applications, 36(29), 18161–18200.

Farangi, M., & Zabbah, I. (2023). Intelligent scoring in an English reading comprehension course using artificial neural networks and neuro-fuzzy systems. Teaching English as a Second Language Quarterly, 42(4), 1–21.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94, 335–343.

Goktepe Yildiz, S., & Goktepe Korpeoglu, S. (2023). Prediction of students' perceptions of problem-solving skills with a neuro-fuzzy model and hierarchical regression method: A quantitative study. Education and Information Technologies, 28(7), 8879–8917.

Haq, I. U., Mohiz, A., Spadafora, G., Palermo, A. M., Bilotta, E., Liccardo, A., & Lamonaca, F. (2025). The role of network of extended reality-enabled laboratories in enhancing STEM education: Bridging theory and practice in the digital classroom. In International Workshop IFToMM for Sustainable Development Goals (pp. 362–372). Cham: Springer.

Haq, I. U., Mohiz, A., Spadafora, G., Palermo, A. M., Bilotta, E., Felicetti, F., & Lamonaca, F. (2025). AI in the network of extended reality-enabled laboratories for STEM education: Current applications and future potential for adaptive learning. In International Workshop IFToMM for Sustainable Development Goals (pp. 373–382). Cham: Springer.

Iatrellis, O., Kotsiantis, S., & Pintelas, P. (2021). Predicting students' graduation time using machine learning techniques: A case study. Education and Information Technologies, 26, 69–88.

Li, Q., Liu, M., Zhao, M., & Tang, X. (2022). Deep learning applications in predicting student performance: A case study in higher education. Educational Data Mining Journal, 14(1), 22–39.

Li, S., Wang, C., & Wang, Y. (2024). Fuzzy evaluation model for physical education teaching methods in colleges and universities using artificial intelligence. Scientific Reports, 14(1), 4788.

Mahmood, Y., et al. (2024). Optimizing natural gas pipeline risk assessment using hybrid fuzzy Bayesian networks and expert elicitation for effective decision-making strategies. Gas Science and Engineering, 125, 205283.

Mehdi, S., & Nachouki, M. (2022). Predicting students' grade point average using adaptive neuro-fuzzy inference systems (ANFIS). International Journal of Artificial Intelligence in Education, 32, 193–215.

Mhady, A. A., Budayan, C., & Gurgun, A. P. (2024). Estimate-at-completion (EAC) prediction using Archimedes optimization with adaptive fuzzy and neural networks. Automation in Construction, 166, 105653.

Mohd Nor, N. H., Mansor, N. I., & Hasim, N. A. (2024). Artificial neural networks: A new frontier in dental tissue regeneration. Tissue Engineering Part B: Reviews. (In press).

Mumali, F., & Kalkowska, J. (2024). Intelligent support in manufacturing process selection based on artificial neural networks, fuzzy logic, and genetic algorithms: Current state and future perspectives. Computers & Industrial Engineering, 110272.

Qian, J., et al. (2024). Optimization of teacher evaluation indicator system based on fuzzy-DEMATEL-BP. Heliyon, 10(13). (Add article number when available.)

Ramesh, B., & Lakshmanna, K. (2024). A novel early detection and prevention of coronary heart disease framework using hybrid deep learning model and neural fuzzy inference system. IEEE Access, 12, 26683–26695.

Saa, A., Al-Emran, M., & Shaalan, K. (2020). An artificial neural network model for predicting student performance in a private university in the UAE. International Journal of Emerging Technologies in Learning (iJET), 15(15), 4–20.

Saville, J. D., & Foster, L. L. (2021). Does technology self-efficacy influence the effect of training presentation mode on training self-efficacy? Computers in Human Behavior Reports, 4, 100124.

Sharafi, M., et al. (2024). Development of long short-term memory along with differential optimization and neural networks for coagulant dosage prediction in water treatment plant. Journal of Water Process Engineering, 65, 105784.

Son, J., & Fujita, H. (2019). A MANFIS-S model for predicting university students' future performance. Knowledge-Based Systems, 178, 151–166.

Tomasevic, N., Gasevic, D., & Loughin, T. M. (2020). Impact of data preprocessing on the performance of predictive models for early identification of students at risk of failing a course. Computers & Education, 144, 103706.

Uskov, V. L., Bakken, J. P., & Byerly, A. (2019). Machine learning in STEM education: Research and applications. Advances in Intelligent Systems and Computing, 950, 225–241.

Xie, B. (2024). An integrated framework for spherical fuzzy MAGDM and applications to English blended teaching quality evaluation. Journal of Intelligent & Fuzzy Systems. (Preprint), 1–17.

Zerouali, B., et al. (2024). Enhancing deep learning-based slope stability classification using a novel metaheuristic optimization algorithm for feature selection. Scientific Reports, 14(1), 21812.