Hybrid Ensemble Learning Model for Measuring the Creativity of Iraqi Gifted Students

Zainab Khyioon Abdalrdha1[(1*)] | **Rawaa waleed Abd Al wahab** [(2)] | **Saad Ahmed Dheyab** [(3)] | **Rawaa Ibrahim Esa**[(4)]

[1*] *Department of Computer Science, College of Education, Mustansiriyah University, Baghdad, Iraq*

[2] *Baghdad School for Gifted Students Authority for the Care of the Gifted, Baghdad, Iraq*

[3] *Baghdad School for Gifted Students Authority for the Care of the Gifted, Baghdad, Iraq*

[4] *Department of Computer Science, College of Basic Education, Mustansiriyah University/ Baghdad, Iraq*

Email Corresponding Author*: zainabkhyioon83@uomustansiriyah.edu.iq

dr.rawaa72@gmail.com

saad.theyab@uoitc.edu.iq

rawaaa_e.edbs@uomustansiriyah.edu.iq

Abstract:

Creativity is the defining trait of giftedness and an essential skill in the age of artificial intelligence (AI). This research presents a hybrid ensemble-learning model to differentiate the creativity profiles of gifted human students and an AI. Six domains of creativity (fluency, flexibility, originality, elaboration, sensitivity, and redefinition) were visualized as the unique features used to build 12,000 data points (24,000 samples, 12,000 responses from gifted human students, and the rest from the generative AI) with identical prompts and completed as the responses to a creativity scale instrument administered to Iraqi gifted students. Six classifiers (decision tree, random forest, K-nearest neighbor, support vector machines, ANN, and

hybrid ensemble model with an eXtreme Gradient Boosting (XGBoost) meta-learner) were tested, and the hybrid ensemble showed the best test accuracy (96.75%), followed by SVM (96.73%).
The cross-validation stability was higher with SVM (mean = 96.69%, SD = 0.0014). MANOVA showed statistically significant multivariate behavior ($p < 0.001$), with gifted students showing significant superiority against the AI in all six creativity aspects. Feature importance weightings showed sensitivity (0.414) and elaboration (0. 191) are the strongest classifiers in differentiating human and AI responses, while discriminating effect sizes for sensitivity (1.464) and originality (1.359) show major differences. AI responses were rated equally high to gifted students' responses in fluency and elaboration but ranked lower in originality, sensitivity, and redefinition. Bottom line, this research is an able and explainable model for scale-based differentiation between human and AI creativity.
Keywords: Gifted Students; Hybrid Ensemble Learning; Machine Learning; Creativity Measurement; Feature Importance; Iraqi Education; XGBoost.

1. Introduction

Creativity is one of the main elements of giftedness, and it is becoming even more important as we move into an AI-driven era. Actually, while education systems worldwide are constantly searching for and training creative potential, mainly among gifted students, many Arab countries still face obstacles when it comes to the ability to measure creativity and identify gifted individuals [1]. Also, the recent breakthroughs in AI have only fueled the desire to explore creativity further, as some research points out, for example, that AI models like GPT-4 are capable of scoring extremely high on creativity scales [2]. Some of these models have even been able to outperform humans in tasks requiring divergent thinking [3]. Besides, creativity assessment and gifted identification systems that depend on machine learning-based approaches have been finding A lot more reliable, scalable, and accurate ways to operate [4, 5]. However, schools are still lacking effective instruments that can accurately differentiate AI-generated creativity from authentic student creativity. A systematic review recently observed that the number of studies on this topic

is increasing rapidly and emphasized the critical need for well-designed, empirically validated assessment systems [6]. This paper intends to introduce and validate a hybrid ensemble learning system that separates human creativity from AI-generated creativity. One of the main modes of analysis is the comparison at the MANOVA level of Iraqi gifted students' answers and AI-generated responses on various creativity dimensions; feature importance analysis is used to pick the most important dimensions, and finally, creativity profiles are developed per the data. Apart from analyzing recent references (2020-2026), the a paper also details dataset development and model design; discusses classification results, statistical analyses, feature importance, educational implications, and major directions for future research.

2. Related Work

Research on combining creativity measurement, machine learning, and human-AI comparison is a great topic for exploration these days as the main drivers of progress in this field include the successful development of ensemble learning techniques and the unleashing of generative AI models to the public. In this review, 2020-2026 related papers are summarized in three main aspects: (a) machine learning methods for creativity evaluation, (b) use of AI for finding gifted students, and (c) more directly human-AI creativity comparisons.

2.1 Machine Learning Approaches to Creativity Assessment

Early research focused mainly on whether deep learning systems could generate and assess creative work. Tests revealed that even though deep learning models can generate novel ideas, they are still limited to the patterns in their training data and lack key human creative abilities like contextual reasoning and redefining problems [7]. Different automated methods for assessing creativity have been introduced, like the Divergent Association Task (DAT) [9] and the Deep Creativity setup [10], which have demonstrated that AI can successfully evaluate different dimensions of creativity, including novelty, value, and surprise. Same here, a deep-learning-based creativity measurement system proved to be on par with traditional assessment techniques with reliability and validity [1]. The

arrival of large language models (LLMs) in 2023 gave a big boost to the studies on the comparative creativity of humans and AI. It was revealed that , on average, AI performance in creative-thinking tasks can be as good as humans', although highly creative individuals can still outperform AI [11]. Besides that, there was a discussion about the issue of the increasing similarity between AI-produced and human creative outputs as well as its effects on creativity evaluation [12]. Other experimental works already showed that highly sophisticated AI systems on some creative tasks can even outperform humans [13]. But, the human-AI partnership is always the one producing the best results [14]. Large-scale research results as of 2025 revealed that in general, there is no major difference between human and AI creativity, but human-AI cooperation leads to the most effective creative outputs [2]. Besides, some researchers discovered that in general AI-generated ideas are less varied and more focused than human ideas [15] and that human creativity still has a stronger association with cognitive exploration, active thinking, and creative flow than AI-assisted creativity [16].

2.2 AI-Based Identification of Gifted Students

In study [3], the authors implemented machine-learning classifiers to merge data from academic and socio-emotional rating scales to identify gifted individuals. They found that integrating different sources of assessment led to better classification accuracy. Like the current research, the authors integrated results of a questionnaire, teacher ratings, and six creativity aspects from TTCT. In Iraqi and Arab education, artificial intelligence (AI) instructional and assessment tools have a significant positive impact on student achievement, although a regional study [4] came up with the suggestion that culturally appropriate gifted identification systems are necessary and that Western instruments should not be used exclusively. Another justification for utilizing context-specific Iraqi gifted child data in this paper. Besides, the paper [17] revealed that Truth is generative AI technology may increase gifted students' creativity and critical thinking capacity. Still, the authors emphasized the requirement for validated methods of differentiating student creativity from AI-generated works.

2.3 Human versus AI Creativity Comparison Studies

Several studies on verbal creativity gave quite similar results. The authors of [18-19] are convinced that large language models surpassed humans on combined verbal creativity tests. GPT-4 scores in the top percentile for many Torrance characteristics of creative thinking, per [6]. More thorough assessments and theoretical examination [20] have revealed the quick progress of AI creativity and the continuing limitations of AI when compared to human creative cognition. Besides, research [21] has demonstrated that the main differentiating factors of human and AI-generated content are distinctiveness and style. Studies show that the emphasis is moving from imagining machine creativity to testing human-AI creative differences on various levels and explaining them. This work was largely inspired by a gap in the literature, which hardly addressed the full integration of multifaceted psychometric creativity measures with ensemble machine learning techniques to identify and interpret the creativity of gifted students versus AI across several creativity dimensions.

2.4 Research Gap

In the past, no one has mixed different creativity aspects based on TTCT, data from questionnaires, students' ratings by teachers, a stacking-type hybrid ensemble model with XGBoost as a meta-learner, and MANOVA for analysis. Besides, our understanding of Iraqi gifted students is very limited. Because of this, the current study offers a hybrid ensemble learning approach that compares and distinguishes human and AI creativity, Because of this addressing both the methodological and contextual gaps.

3. Proposed Methodology

The new approach is a pipeline of six stages shown in Figure 1, which involves collecting data, preprocessing data, constructing a hybrid ensemble model, predicting creativity score, training and evaluating the model, and finally assessing performance with standard metrics.

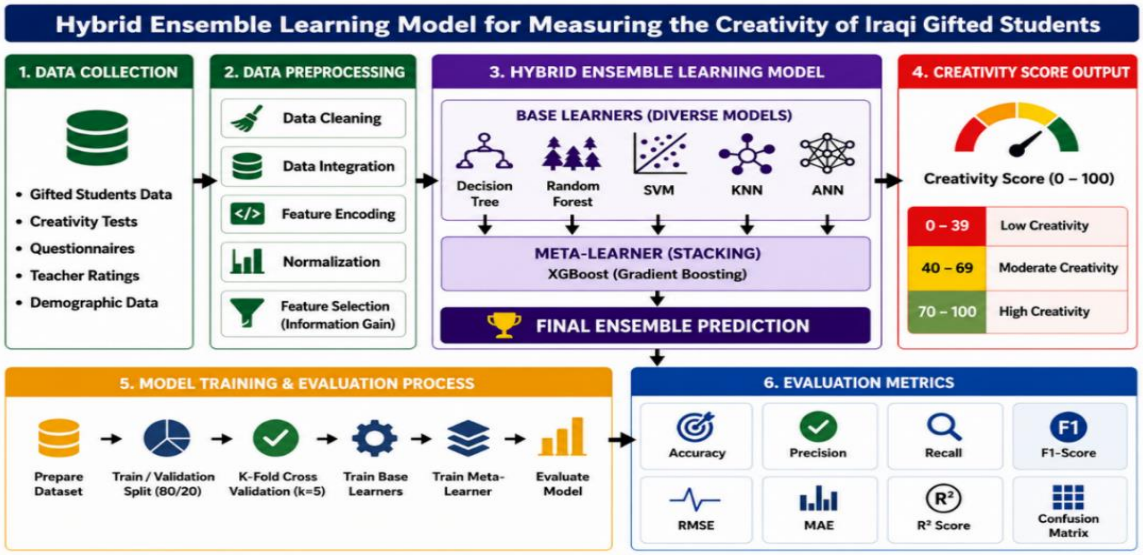


Figure 1. An Iraqi gifted students' creativity measurement hybrid ensemble learning model architecture.

3.1 Dataset Description

The authors of this research paper have used authentic information for their study of (540) Iraqi gifted students, among whom 134 students were from Baghdad. Creativity has been measured using a standard creativity instrument across six different aspects. To get comparable AI-generated answers, we also gave the same tool and scoring system to a generative AI program. This made sure that the comparison between human and AI creativity was done on the same tasks and evaluation criteria. The created dataset had 24,000 balanced samples (12,000 human and 12,000 AI responses). Seven features of input were present in each sample: the six creativity traits (fluency, flexibility, originality, elaboration, sensitivity, and redefinition) with other information from the questionnaire scores and teacher ratings. The target variable marked the responses as human (1) or AI (0), and that was used for both training and testing the proposed classification models. Representative examples from the Human-AI Creativity Dataset used in this study, illustrating in Figure 2 five human-generated samples (Source = 1) and five AI-generated samples (Source = 0).

Human Samples (Source = 1)

Student_ID	Fluency	Flexibility	Originality	Elaboration	Sensitivity	Redefinition	Questionnaire	Teacher_Rating	Portfolio	Creativity_Score	Creativity_Level	Source
1	61.52	61.72	62.90	58.56	59.31	59.17	50.43	57.18	65.63	60.15	2	1
2	61.62	69.04	64.53	66.00	65.83	67.07	64.46	63.40	65.05	64.99	2	1
3	84.05	82.14	80.98	85.57	81.19	79.17	81.46	81.89	83.64	83.30	0	1
4	71.23	72.26	74.97	74.57	70.57	71.54	66.09	66.47	60.91	70.51	0	1
5	62.99	64.60	65.56	61.69	63.05	60.29	66.86	58.82	70.39	64.08	2	1

AI Samples (Source = 0)

Student_ID	Fluency	Flexibility	Originality	Elaboration	Sensitivity	Redefinition	Questionnaire	Teacher_Rating	Portfolio	Creativity_Score	Creativity_Level	Source
1	70.21	70.78	66.53	70.02	62.02	66.64	68.08	71.79	72.11	67.91	2	0
2	66.16	60.84	55.22	64.55	55.65	59.30	60.43	61.95	70.07	60.35	2	0
3	65.49	66.25	61.56	66.04	61.09	60.14	65.53	66.46	73.89	64.11	2	0
4	64.36	59.27	56.13	58.79	52.06	56.11	62.23	63.93	62.36	58.66	1	0
5	59.17	58.69	55.72	61.29	49.08	54.26	57.58	58.45	62.44	58.09	1	0

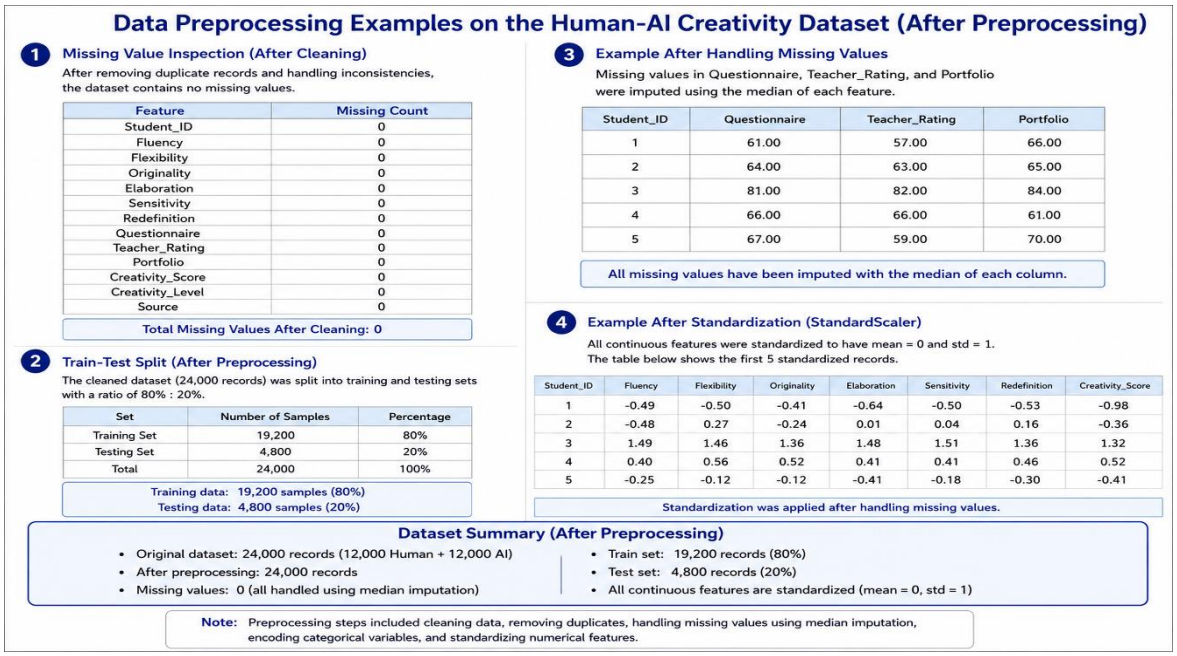
Figure 2. Examples from the Human-AI Creativity Dataset.

3.2 Data Preprocessing

As shown in Stage 2 of Figure 1, the first five steps of the pre-processing pipeline are performed sequentially and independently before the model is trained [22-23]:

- Data Cleaning: identifying missing values, eliminating duplicate records, and locating outliers in all eight features.
- Data Integration: consolidation of creativity test scores, questionnaire responses, and teacher ratings into a unified tabular dataset.
- Feature Encoding: encoding of the categorical group label (Human/AI) into a binary numerical format suitable for classification.
- Normalization: standardizing (z-score normalization) all continuous features to make different variables measured on different scales comparable. This is very important for distance-based algorithms (KNN, SVM) and also gradient-based algorithms (ANN).
- Feature Selection: usage of an information gain scale to measure the potential of each of the eight features in the classification task before model training.

Figure 3 depicts the multiple preprocessing steps carried out on the Human-AI Creativity Dataset. The first step was to check for missing values, followed by median imputation, train-test splitting (80% training and 20% testing), and finally feature standardization using StandardScaler. These are some of the ways in which the data were prepared for model building.



Data Preprocessing Examples on the Human-AI Creativity Dataset (After Preprocessing)

1 Missing Value Inspection (After Cleaning)

After removing duplicate records and handling inconsistencies, the dataset contains no missing values.

Feature	Missing Count
Student_ID	0
Fluency	0
Flexibility	0
Originality	0
Elaboration	0
Sensitivity	0
Redefinition	0
Questionnaire	0
Teacher_Rating	0
Portfolio	0
Creativity_Score	0
Creativity_Level	0
Source	0

Total Missing Values After Cleaning: 0

2 Train-Test Split (After Preprocessing)

The cleaned dataset (24,000 records) was split into training and testing sets with a ratio of 80% : 20%.

Set	Number of Samples	Percentage
Training Set	19,200	80%
Testing Set	4,800	20%
Total	24,000	100%

Training data: 19,200 samples (80%)
Testing data: 4,800 samples (20%)

3 Example After Handling Missing Values

Missing values in Questionnaire, Teacher_Rating, and Portfolio were imputed using the median of each feature.

Student_ID	Questionnaire	Teacher_Rating	Portfolio
1	61.00	57.00	66.00
2	64.00	63.00	65.00
3	81.00	82.00	84.00
4	66.00	66.00	61.00
5	67.00	59.00	70.00

All missing values have been imputed with the median of each column.

4 Example After Standardization (StandardScaler)

All continuous features were standardized to have mean = 0 and std = 1. The table below shows the first 5 standardized records.

Student_ID	Fluency	Flexibility	Originality	Elaboration	Sensitivity	Redefinition	Creativity_Score
1	-0.49	-0.50	-0.41	-0.64	-0.50	-0.53	-0.98
2	-0.48	0.27	-0.24	0.01	0.04	0.16	-0.36
3	1.49	1.46	1.36	1.48	1.51	1.36	1.32
4	0.40	0.56	0.52	0.41	0.41	0.46	0.52
5	-0.25	-0.12	-0.12	-0.41	-0.18	-0.30	-0.41

Standardization was applied after handling missing values.

Dataset Summary (After Preprocessing)

- Original dataset: 24,000 records (12,000 Human + 12,000 AI)
- After preprocessing: 24,000 records
- Missing values: 0 (all handled using median imputation)
- Train set: 19,200 records (80%)
- Test set: 4,800 records (20%)
- All continuous features are standardized (mean = 0, std = 1)

Note: Preprocessing steps included cleaning data, removing duplicates, handling missing values using median imputation, encoding categorical variables, and standardizing numerical features.

Figure 3. Data Preprocessing Example on the Dataset.

3.3 Hybrid Ensemble Learning Model Architecture

The main element of the proposed approach, shown in Stage 3 of Figure 1, is a two-tier stacking system containing five very different base learners and just one meta-learner.

3.3.1 Base Learners

Five base classifiers that we picked for their complementary inductive biases were trained on the preprocessed feature set:

• Decision Tree: Contains clear, comprehensible rules and non-linear feature interactions [24].

• Random Forest: Multiple decision trees reduce overfitting risk through bootstrap aggregation (bagging) and improve robustness by randomly selecting features at each split point [25].
• SVM uses a perfect dividing hyperplane to optimize the margin between human and AI creative profiles, which is effective in the high-dimensional, continuous creativity feature space [25].
• K-Nearest Neighbors (KNN) assigns a sample to the class most common among its nearest Neighbors, effectively capturing local similarity patterns in the feature space [26].
• ANN: implements a multi-layer perceptron which is capable of modelling complex, non-linear interactions of eight creativity-related features [27].

3.3.2 Meta-Learner (Stacking)

In the third stage of the model suggested here, XGBoost acts as a meta-learner [28] within a stacking ensemble setting. The class-probability output of the base classifiers forms feature inputs for XGBoost for finding the optimum combination of their predictions. This simply means that the method individually corrects the mistakes of the base models and produces the final ensemble prediction. The stacking-based system, in other words, the hybrid ensemble model is the subject of the present study.

3.4 Creativity Score Output

The output of the hybrid ensemble model is a creativity score between 0 and 100 based on Stage 4 of Figure 1. To facilitate their use in real educational settings, these are broken down into three qualitative bands:

- 0-39: Low Creativity
- 40-69: Moderate Creativity
- 70-100: High Creativity

The classification system with three levels offers an easy and understandable way to explain results of creativity assessment to different stakeholders like teachers, parents, and gifted education administrators.

3.5 Model Training and Evaluation Process

Consulting Stage 5 in Figure 1, the steps for model training and evaluation can be demonstrated:

1. Dataset Preparation: The completely cleaned, merged, and normalized dataset consisting of 24,000 examples is set for model use.
2. Train/Validation Split: The dataset will be divided with 80/20 ratio, the stratification being applied to keep the same balanced distribution of classes (12,000 samples human and 12,000 samples AI) in proportion within each partition as well.
3. K-Fold Cross-Validation (k = 5): a five-fold cross-validation approach is used in the training section to evaluate the model's
stability and ability to generalize before the final evaluation on the separate test set.

1. Dataset Preparation: The completely cleaned, merged, and normalized dataset consisting of 24,000 examples is set for model use.
2. Train/Validation Split: The dataset will be divided with 80/20 ratio, the stratification being applied to keep the same balanced distribution of classes (12,000 samples human and 12,000 samples AI) in proportion within each partition as well.
3. K-Fold Cross-Validation (k = 5): a five-fold cross-validation approach is used in the training section to evaluate the model's stability and ability to generalize before the final evaluation on the separate test set.

4. Train Base Learners: Each of the five base classifiers (Decision Tree, Random Forest, SVM, KNN, ANN) is trained independently on the training partition.
5. Train Meta-Learner: the XGBoost meta-learner is trained on the out-of-fold predictions of the base learners, completing the stacking architecture.
6. Evaluate Model: the final Hybrid Ensemble model, along with each individual base learner, is evaluated on the held-out test set using the metrics described in Section 3.6.

3.6 Evaluation Metrics

As shown in Stage 6 of Figure 1, model performance is assessed using a comprehensive suite of evaluation metrics [29], comprising both classification metrics, accuracy, precision, recall, and F1-score; supplementary regression diagnostic metrics, namely Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [30], and the coefficient of determination (R^2 Score); and the confusion matrix. Classification metrics serve as the primary means of comparing the six models (decision tree, random forest, KNN, SVM, ANN, and hybrid ensemble), whereas the confusion matrix provides an in-depth analysis of classification errors for each class. These are statistical evaluation metrics used to assess machine-learning models. They are not educational measures but quantitative indicators used to evaluate model performance, as illustrated in Table 1, Summary of Machine Learning Evaluation Metrics.

Table 1. Summary of Machine Learning Evaluation Metrics.

Metric	Description	Equations	Interpretation
Accuracy	**Determines the fraction of total observations that were classified correctly.**	$\text{Acc.} = (T_P + T_N)/(T_P + T_N + F_P + F_N)$	**The nearer the value is to 1, the more accurate it is.**
Precision	**Determines the fraction of correctly predicted positive cases among all predicted positives.**	$\text{Pre.} = T_P/(T_P + F_P)$	**The more the values, the less false positives.**
Recall	**Determines how well the model detects all actual positive cases.**	$\text{Rec.} = T_P/(T_P + F_N)$	**The more the values, the less missed positive cases.**
F1-Score	**The middle value in the range of Precision and Recall.**	$\text{F1.} = 2 \times (\text{Pre.} \times \text{Rec.})/(\text{Pre.} + \text{Rec.})$	**Good for imbalanced datasets.**
MAE	**Measures the mean discrepancy between true values and the ones**	$\text{MAE} = (1/n)\Sigma\lvert yi - \hat{y}i\rvert$	**Better performance corresponds to lower values.**

	predicted.		
RMSE	**Measures the extent of the mean-squared prediction errors in square units.**	$\mathbf{RMSE} = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} (\mathbf{yi} - \hat{\mathbf{y}}\mathbf{i})^2$	**Higher prediction accuracy means lower values.**
R^2 Score	**Represents the fraction of total variance that the model captures.**	$\mathbf{R^2\ Score} = \frac{\sum_{i=1}^{n} (\mathbf{Xi} - \mathbf{Yi})^2}{\sum_{i=1}^{n} (\bar{Y} - Y_i)^2}$	**The closer the values are to 1, the better the explanatory power.**
Confusion Matrix	**A chart that displays the comparison between actual and predicted classifications.**	T_P, T_N, F_P, F_N	**Classification metrics are derived from these.**

3.7 Statistical Comparison Framework

To examine the differences between the creative profiles of humans and AIs, a MANOVA [31] was run simultaneously across six creativity dimensions. Four multivariate test statistics were employed: Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace and Roy's Greatest Root. Besides, independent-samples t-tests were done for each creativity dimension as well as for the overall creativity score. The study presented group means, standard deviations, t-values, p-values, and Cohen's d effect sizes to measure the size of differences between the human and AI groups.

3.8 Feature Importance Analysis

The hybrid ensemble model was subjected to feature importance analysis to determine the variables that first differentiated human and AI creativity profiles. This study looked at the relative importance of all the features used by the model like six creativity dimensions, questionnaire scores, and teacher ratings in classifying the human and AI profiles.

4. Results and Discussion

4.1 Classification Performance

Table 2 shows the final held-out test set performance comparison of all six models.

Table 2. All six models' final categorization statistics.

Model	Accuracy	Precision	Recall	F1-Score
SVM	96.73%	96.74%	96.73%	96.73%
Random Forest	96.58%	96.58%	96.58%	96.58%
KNN	96.15%	96.21%	96.15%	96.14%
ANN	95.73%	95.73%	95.73%	95.73%
Decision Tree	93.52%	93.53%	93.52%	93.52%
Hybrid Ensemble	**96.75%**	**96.75%**	**96.75%**	**96.75%**

Using an XGBoost meta-learner to combine the prediction powers of five base learners, the model detects multiple elements of a creativity dataset and becomes less dependent on a single classifier. The decision tree had the lowest accuracy (93. 52%), which might be due to overfitting and lack of capability to handle complex feature interdependencies.

4.2. Cross-Validation Results

Cross-validation by 5-fold of the hybrid ensemble was very much in line, showing high accuracies around 96.15 %-6.85 %, with an average of 6.69% and a very low SD of 0. 0014. The hybrid ensemble model was demonstrated by the sum total of these three indicators to be a strong, stable, and consistent generalizer for the real-world task of Creativity assessment and Gifted Education in Iraq.

4.3. Classification Report and Confusion Matrix Analysis

Table 3 and Figure 4 summarizes the detailed categorization report's balanced performance across classes.

Table 3. Hybrid ensemble model class-wise assessment.

Class	Precision	Recall	F1-Score	Support
AI (0)	96.00%	97.00%	97.00%	12000
Human (1)	97.00%	96.00%	97.00%	12000
Accuracy			97.00%	24000
Macro Avg	97.00%	97.00%	97.00%	24000
Weighted Avg	97.00%	97.00%	97.00%	24000

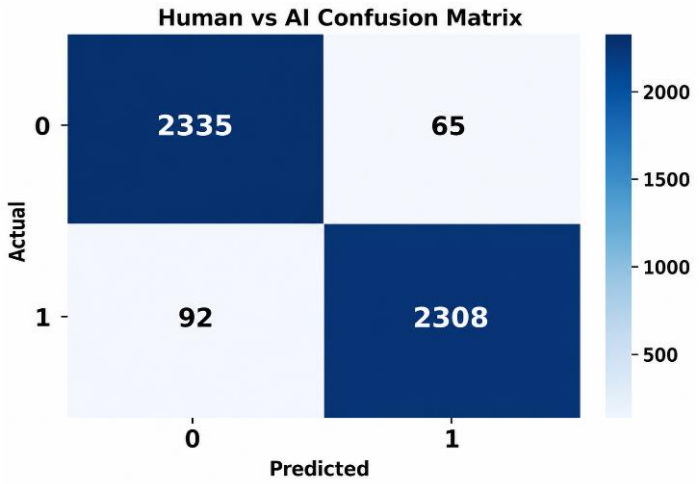


Figure 4: Confusion matrix for the categorization job between humans and AI.

The confusion matrix reveals that the hybrid ensemble model reached an accuracy of 97% on a dataset with 24,000 samples and it made only a few errors. The almost equal division of human and AI classification errors shows that the model is not biased towards one or the other. The model's high precision, recall and F1-score results indicate not only its trustworthiness but also its capability to separate human from A creative profiles.

4.4 Feature Importance

Feature importance analysis showed the order of the seven features used for differentiating the human creativity profile from the AI creativity profile, as presented in Table 4 and Figure 5.

Table 4. Component hybrid ensemble model feature relevance ranking.

Rank	Feature	Importance
1	Sensitivity	0.4139
2	Elaboration	0.1912
3	Originality	0.1260
4	Flexibility	0.0999
5	Redefinition	0.0699
6	Questionnaire	~0.052
7	Teacher Rating	~0.047

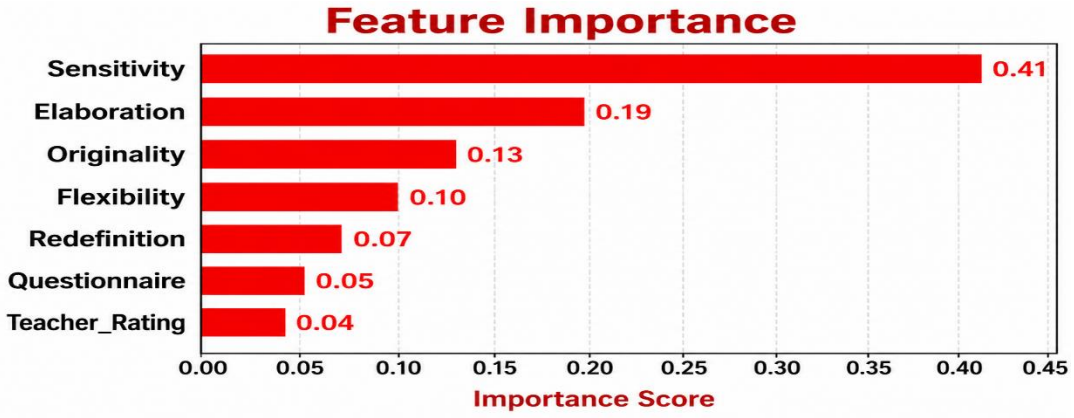


Figure 5. Feature importance scores for the seven predictors used in the classification model.

Then again, sensitivity was not only the most influential feature, but it also had the biggest effect size when we compared humans and AI. This means that sensitivity is the main reason that differentiates human and AI creativity profiles. Same thing, originality showed a considerable discriminative effect,

whereas elaboration, although it had a smaller effect size between the groups, was a very important feature for classification.

4.5 Human vs. AI Creativity Profile Comparison

Table 5 shows the standardized creativity scores (based on a scale from 0 to 100) for the Human and AI groups in each of the six creativity dimensions. In addition, Figures 6 and 7 display the comparison of average creativity scores between the human and AI groups along all six dimensions, as well as a composite overall creativity score.

Table 5. Descriptive statistics and independent-samples t-test outcomes for human and AI comparisons in creativity dimensions.

Dimension	Human Mean	AI Mean	Human SD	AI SD	t-value	p-value	Cohen's d	Higher
Fluency	72.026	68.012	10.142	5.379	38.301	< .001	0.494	Human
Flexibility	71.985	67.007	10.389	5.367	46.630	< .001	0.602	Human
Originality	72.274	61.007	10.420	5.375	105.258	< .001	1.359	Human
Elaboration	72.031	67.999	10.714	5.359	36.877	< .001	0.476	Human
Sensitivity	72.051	59.993	10.339	5.371	113.370	< .001	1.464	Human
Redefinition	72.027	60.999	10.409	5.383	103.087	< .001	1.331	Human
Creativity Score	72.038	64.662	10.052	5.162	71.497	< .001	0.923	Human

Humans, as a group, obtained higher average scores than the AI group in all six creative dimensions plus overall creativity score. The differences were highly significant statistically ($p < .001$). Sensitivity ($d = 1.464$), originality ($d = 1.359$), and redefinition ($d = 1.331$) were the aspects in which differences were the biggest, implying that these three dimensions are the main factors that differentiate human creativity from that of AI. Meanwhile, fluency ($d = 0.494$) and elaboration ($d = 0.476$) were examples of dimensions that had less difference; Still, human participants still

outperformed AI. The results, in general, imply that human creativity is defined most Really by sensitivity, originality, and redefinition.

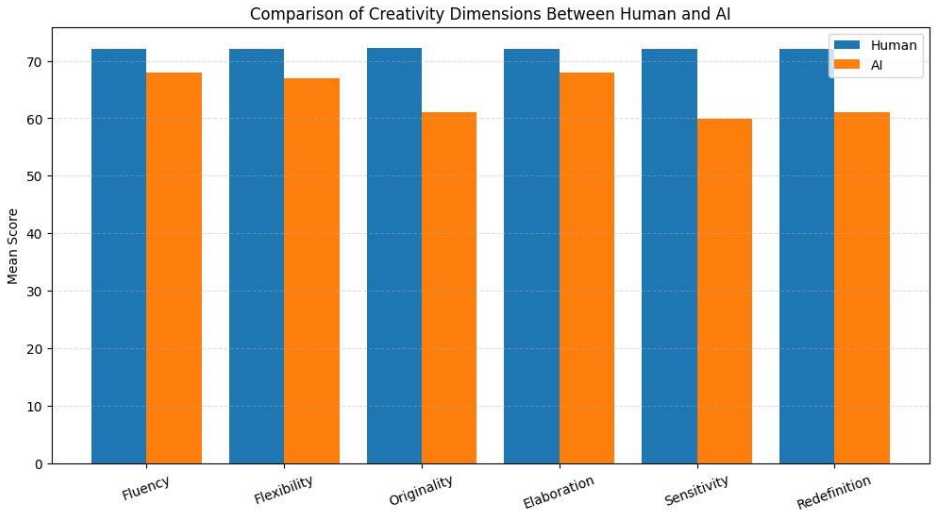


Figure 6. Comparing the average scores of six creativity dimensions between the human and AI groups

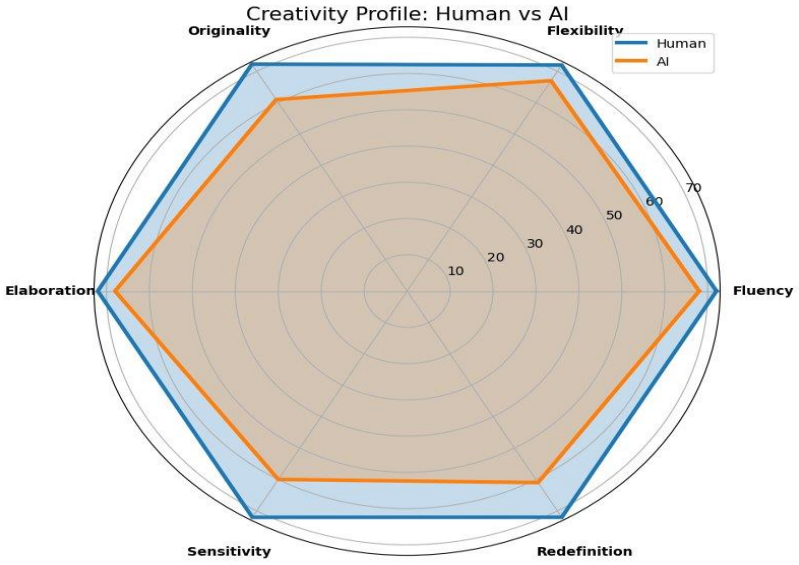


Figure 7. Radar chart illustrating the creativity profile of humans (Iraqi gifted students) and AI across the six dimensions.

The creativity profiles of the two groups, human and AI, were measured and compared on the radar graph as shown in Figure 7. Across all six dimensions

of creativity, the human profile was distinctly higher than the AI one. In fact, on the creativity dimensions of sensitivity, originality, and redefinition, the differences between the two profiles were very pronounced. But the two profiles were still quite near one another on fluency and elaboration, meaning that the performance differences in these dimensions were smaller.

4.6 Multivariate Analysis of Variance (MANOVA)

According to the MANOVA results in Table 6, the group category (Human vs. AI) had a very significant multivariate influence over the six creativity dimensions. The four different multivariate tests, Wilks' Lambda (0.2357), Pillai's Trace (0.7643), Hotelling-Lawley Trace (3.2425), and Roy's Greatest Root (3.2425), were all statistically significant at $p < 001$. This way consistently demonstrates that the human and AI groups differ when all six creativity dimensions are taken into account simultaneously.

Table 6. MANOVA test statistics for the human vs. AI group effect across six creativity dimensions.

Statistic	Value	F-value	df1	df2	p-value
Wilks' Lambda	0.2357	12966.37	6	23993	< .001
Pillai's Trace	0.7643	12966.37	6	23993	< .001
Hotelling-Lawley Trace	3.2425	12966.37	6	23993	< .001
Roy's Greatest Root	3.2425	12966.37	6	23993	< .001

Both a very low Wilks' Lambda figure (0.2357) and a very high Pillai's Trace figure (0.7643) suggest the two groups are distinctly separated on a multivariate level. Our interpretation of the data is based on the extremely large F-value (12966.37), which implies that these differences were very unlikely to have happened at random. That means the null hypothesis that there is no multivariate difference between human and AI creativity profiles is rejected. These data signify that the creative profiles of Iraqi gifted students and of AI-generated samples are at a multivariate level highly

different, which not only strengthens the findings of the univariate analyses but also substantiates the capability of the proposed Hybrid Ensemble model in differentiating human and AI creativity profiles.

4.7 Synthesis of Findings

By integrating the results of the classification, the determination of feature importance, and the statistical comparisons, a very consistent set of conclusions is derived:

1. The Hybrid Ensemble model can classify human and AI creativity profiles almost at the state-of-the-art accuracy level (96.73%), and it is very dependable when tested on different data sets (standard deviation = 0.0014).
2. The performance of the classification model is mostly dependent on the dimensions of sensitivity, elaboration, and originality. These are also the dimensions that most strongly and statistically differentiate the human and AI groups, showing that both machine learning and statistical analyses are in agreement.
3. Iraqi gifted students have a unique creativity profile with mostly high levels of sensitivity, originality, and redefinition abilities compared to the AI benchmarks. They also have smaller, yet statistically significant, advantages in fluency, flexibility, and elaboration.
4. MANOVA confirms that this is in fact a genuine multivariate distinction and not a result of testing multiple dimensions separately.

4.8 Comparison with Previous Studies

To confirm the results of the proposed model, human-versus-AI creativity studies were compared. The main differences in the research objectives, datasets, methods, and results of these studies are summarized in Table.7. The comparison shows that the proposed Hybrid Ensemble Learning model is effective in differentiating creativity profiles between gifted students and AI.

Table 7. Comparison between the Present Study and Previous Studies

Study	Dataset	Method	Main Result
[12]	Human and AI creative tasks	Comparative analysis	Human creativity remained superior in originality and novelty
[17]	No experimental dataset	Educational review	No quantitative results reported
[20]	No experimental dataset	Conceptual analysis	AI showed limitations in originality and surprise
[21]	No experimental dataset	Survey review	AI struggled with abstract reasoning and originality
[22]	500,000 essays	Random Forest, BERT	Best Accuracy = 90.88%
Present Study	24,000 records	Hybrid Ensemble (XGBoost)	**Accuracy = 96.75%**

5. Conclusion

The research created and confirmed an AI-human hybrid ensemble-learning model that can identify the creativity profiles of Iraqi gifted students as well as AI-generated answers, representing six creativity dimensions: fluency, flexibility, originality, elaboration, sensitivity, and redefinition. The model was a two-level stacking architecture: five base learners (decision tree, random forest, SVM, KNN, and ANN) and an XGBoost meta-learner. The model that was suggested got a classification accuracy of 96.75% together with excellent cross-validation reliability (mean = 96.69%, SD = 0.0014) and beat the best individual classifier (SVM, 96.73%) only by a slight margin. Results of MANOVA showed that there are extremely significant differences between creativity profiles of humans and AI (Wilks' Lambda = 0.236, Pillai's Trace = 0.764, $p < .001$), with gifted students being quite a bit superior in all six creativity dimensions. Per the feature importance analysis, sensitivity (41.4%) and elaboration (19.1%) were the most critical factors, whereas the biggest differences between human and AI creativity were in sensitivity ($d = 1.464$), originality ($d = 1.359$), and redefinition ($d = 1.331$). That means that originality, sensitivity, and redefinition are the main features

of human creativity, and the results provide a strong and scalable system for creativity assessment and gifted education in the Iraqi educational system. Identifying and accepting some of the shortcomings of the present research is one way to indicate the future research directions:

- Longitudinal follow-up: Other researchers may consider the possibility of carrying out this kind of work by using the model with time-series data to discover the changes in creativity profiles.
- Dimensional expansion: One way of expanding the assessment of one's creative potential is by including other dimensions of creativity and the use of additional psychometric instruments.
- Cross-architecture generalizability: The capability of the proposed structure to be effective with different AI architectures needs to be ascertained by examining it with creativity profiles derived from various generative AI systems, even those that are multimodal, to reveal its potency.
- Cross-regional validation: Validation of the setup should be done by taking the gifted student populations from different educational and cultural settings so that it can be ascertained whether the findings are capable of being generalized beyond Iraq.

Acknowledgement

The authors would like to thank Al-Mustansiriyah University (www.uomusiriyah.edu.iq),and Baghdad School for Gifted Students Authority for the Care of the Gifted Baghdad-Iraq for its support in the present work.

References

[1] Sung, Y. T., Cheng, H. H., Tseng, H. C., Chang, K. E., & Lin, S. Y. (2022). *Construction and Validation of a Computerized Creativity Assessment Tool with Automated Scoring Based on Deep-Learning Techniques*. Psychology of Aesthetics, Creativity, and the Arts.

[2] Holzner, N., Maier, S., & Feuerriegel, S. (2025). *Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis*. arXiv.

[3] Lee, H., Cho, J., & Walsh, A. (2025). Enhancing Gifted Identification: A Machine Learning Analysis of the HOPE Teacher Rating Scale Responses.

Journal of Advanced Academics.
https://doi.org/10.1177/1932202X251362945
[4] Gifted education in Arab countries: analyses from a learning-resource perspective. (2022). Cogent Education, 9(1).
https://doi.org/10.1080/2331186X.2022.2115620
[5] Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14, 3440. https://doi.org/10.1038/s41598-024-53303-w
[6] Guzik, E. E., Byrge, C., & Gilde, C. (2023). The originality of machines: AI takes the Torrance test. *Journal of Creativity*, 33(3), 100065. https://doi.org/10.1016/j.yjoc.2023.100065.
[7] Basalla, M., Schneider, J., & vom Brocke, J. (2020). *Creativity of Deep Learning: Conceptualization and Assessment*. arXiv:2012.02282.
[8] Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). *Naming unrelated words predicts creativity*. Proceedings of the National Academy of Sciences, 118(25), e2022340118.
[9] Franceschelli, G., & Musolesi, M. (2022). *DeepCreativity: Measuring Creativity with Deep Learning Techniques*. arXiv:2201.06118.
[10] Haase, J., & Hanel, P. H. P. (2023). *Artificial Muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity*. Journal of Creativity, 33(3), 100066.
[11] Acar, S. (2023). *Creativity Assessment, Research, and Practice in the Age of Artificial Intelligence*. Creativity Research Journal, 37(2), 181–187.
[12] Wang, H., Zou, J., Mozer, M., Goyal, A., Lamb, A., Zhang, L., et al. (2024). *Can AI Be as Creative as Humans?* arXiv.
[13] Maltese, A. G., Pelletier, P., & Guichardaz, R. (2024). *Can AI Enhance its Creativity to Beat Humans?* arXiv.
[14] Nave, G., Terwiesch, C., & Meincke, L. (2025). *Human–AI Brainstorming and Creativity Diversity*. Nature Human Behaviour.
[15] Song, P. (2026). *Redefining Creativity Differences for Art Students Who Use Generative AI*. Humanities and Social Sciences Communications.

[16] Altememy, H. A., Mohammed, B. A., Hsony, M. K., Hassan, A. Y., Mazhair, R., Dawood, I. I., & Sharif, H. R. (2023). The influence of the artificial intelligence capabilities of higher education institutions in Iraq on students' academic performance: The role of AI-based technology application as a mediator. Eurasian Journal of Educational Research, 104, 267-282.

[17] Siegle, D. (2023). A Role for ChatGPT and AI in Gifted Education. Gifted Child Today, 46(3), 211-219. https://doi.org/10.1177/10762175231168443.

[18] Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., Bengio, Y., & Jerbi, K. (2026). Divergent creativity in humans and large language models. *Scientific Reports*, 16, 1279. https://doi.org/10.1038/s41598-025-25157-3

[19] Dinu, A., & Florescu, A. M. (2024). An integrated benchmark for verbal creativity testing of LLMs and humans. *Procedia Computer Science*, 246, 2902-2911.

[20] Franceschelli, G., & Musolesi, M. (2024). On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

[21] Ismayilzada, M., Paul, D., Bosselut, A., & van der Plas, L. (2024). Creativity in AI: Progresses and challenges. *arXiv preprint arXiv:2410.17218*.

[22] Decoding AI and Human Authorship: Nuances Revealed Through NLP and Statistical Analysis. *arXiv preprint arXiv:2408.00769*. https://arxiv.org/pdf/2408.00769.

[23] Koukaras, Paraskevas, and Christos Tjortjis. 2025. "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices" AI 6, no. 10: 257. https://doi.org/10.3390/ai6100257.

[24] Abdalrdha, Z. K., Mohsin Al-Bakry, P. D. A., & K. Farhan , P. D. A. (2024). Arabic Crime Tweet Filtering and Prediction Using Machine Learning. Iraqi Journal for Computers and Informatics, 50(1), 73–85. https://doi.org/10.25195/ijci.v50i1.479

[25] Abdalrdha, Z.K., Kadhim, A.A., Kadum, A., Naser, W.A.K. (2025). Enhancing fake news detection via PSO-Optimized Ensemble learning: A

comparative study of SVM, NB, and RF. Ingénierie des Systèmes d'Information, Vol. 30, No. 6, pp. 1629-1638. https://doi.org/10.18280/isi.300621.

[26] Sultana, Zinnia, Ashifatul Ferdousi, Farzana Tasnim, and Lutfun Nahar. "An improved k-nearest neighbor algorithm for pattern classification." *International Journal of Advanced Computer Science and Applications* 13, no. 8 (2022): 760-767. DOI: 10.14569/IJACSA.2022.0130887.

[27] Li, Donghong, and Ziyang Liu. 2024. "Artificial Neural Networks (ANNs) and Machine Learning (ML) Modeling Employee Behavior with Management Towards the Economic Advancement of Workers" Sustainability 16, no. 21: 9516. https://doi.org/10.3390/su16219516.

[28] Al-bakri, Fatima Hasan, Wan Mohd Yaakob Wan Bejuri, Mohamed Nasser Al-Andoli, Raja Rina Raja Ikram, Hui Min Khor, Zulkifli Tahir, and The Alzheimer's Disease Neuroimaging Initiative. 2025. "A Meta-Learning-Based Ensemble Model for Explainable Alzheimer's Disease Diagnosis" Diagnostics 15, no. 13: 1642. https://doi.org/10.3390/diagnostics15131642.

[29] Kim, Chaesu, Haeun Park, and Joshua (Sung) You. 2023. "Ecological Fall Prediction Sensitivity, Specificity, and Accuracy in Patients with Mild Cognitive Impairment at a High Risk of Falls" Sensors 23, no. 15: 6977. https://doi.org/10.3390/s23156977.

[30] Jadon, A., Patil, A., Jadon, S. (2024). A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting. In: Sharma, N., Goje, A.C., Chakrabarti, A., Bruckstein, A.M. (eds) Data Management, Analytics and Innovation. ICDMAI 2024. Lecture Notes in Networks and Systems, vol 998. Springer, Singapore. https://doi.org/10.1007/978-981-97-3245-6_9.

[31] Smith, K. N., Lamb, K. N., & Henson, R. K. (2020). Making Meaning out of MANOVA: The Need for Multivariate Post Hoc Testing in Gifted Education Research. Gifted Child Quarterly, 64(1), 41-55. https://doi.org/10.1177/0016986219890335.

وقائع المؤتمر العلمي السنوي الثالث لكلية التربية الاساسية
في مجال العلوم الصرفة والموسوم:
(مؤتمر كلية التربية الاساسية في مجال العلوم الصرفة)
وتحت شعار
(بالعلوم الصرفة نبني واقعا معرفيا متجددا)
يومي الاحد والاثنين 21-2026/6/22

نموذج تعلم تجميعي هجين لتقييم إبداع الطلبة الموهوبين في العراق

زينب خيون عبدالرضا [1]
قسم علوم الحاسوب، كلية التربية الجامعة المستنصرية بغداد، العراق
07722631145
Email Corresponding Author*: zainabkhyioon83@uomustansiriyah.edu.iq

رواء وليد عبد الوهاب [2]
مدرسة الموهوبين ، هيأة رعاية الموهوبين ، بغداد، العراق
07901642193
evanalabidy@uomustansiriyah.edu.iq

سعد احمد ذياب [3]
مدرسة الموهوبين ، هيأة رعاية الموهوبين ، بغداد، العراق
07716008811
Mohammed.s.alhammash@uomustansiriyah.edu.iq

رواء إبراهيم عيسى [4]
قسم علوم الحاسوب، كلية التربية الأساسية ،الجامعة المستنصرية بغداد، العراق
07711599285
rawaaa_e.edbs@uomustansiriyah.edu.iq

ملاحظة : هل البحث مستل من رسالة ماجستير او اطروحة دكتوراه ؟ كلا

المستخلص البحث:

يُعد الإبداع السمة الجوهرية للموهبة وإحدى المهارات الأساسية في عصر الذكاء الاصطناعي. يقدّم هذا البحث نموذجًا هجينًا للتعلّم التجميعي (Hybrid Ensemble Learning) يهدف إلى التمييز بين أنماط الإبداع لدى الطلبة الموهوبين والذكاء الاصطناعي التوليدي. واعتمدت الدراسة ستة أبعاد للإبداع، وهي: الطلاقة، والمرونة، والأصالة، والإثراء (التفصيل)، والحساسية للمشكلات، وإعادة التعريف، بوصفها خصائص أساسية لبناء قاعدة بيانات تضم (24,000)عينة، منها (12,000)استجابة لطلبة موهوبين و (12,000)استجابة مولدة بواسطة الذكاء الاصطناعي، وذلك باستخدام المتغيرات نفسها والاستجابة لأداة قياس الإبداع المطبقة على الطلبة الموهوبين العراقيين. تم اختبار ستة مصنفات للتعلّم الآلي، وهي: شجرة القرار، والغابة العشوائية، وخوارزمية أقرب الجيران(KNN) ، وآلات المتجهات الداعمة(SVM)، والشبكات العصبية الاصطناعية(ANN) ، بالإضافة إلى نموذج تعلّم تجميعي هجين يعتمد على XGBoost (extreme Gradient Boosting)بوصفه متعلّمًا فوقيًا(Meta-Learner) . أظهرت النتائج أن النموذج الهجين حقق أعلى دقة تصنيف في بيانات الاختبار بلغت 96.75%، متفوقًا على نموذج SVM الذي حقق دقة بلغت 96.73%. كما أظهر نموذج SVM استقرارًا أعلى في التحقق المتقاطع بمتوسط دقة بلغ 96.69% وانحراف معياري مقداره 0.0014 . وأظهرت نتائج تحليل التباين المتعدد (MANOVA) وجود فروق ذات دلالة إحصائية عالية ($p < 0.001$) بين استجابات الطلبة الموهوبين واستجابات الذكاء الاصطناعي في جميع أبعاد الإبداع الستة. كما كشفت أوزان أهمية الخصائص أن الحساسية للمشكلات (0.414)والإثراء (التفصيل (0.191) (كانتا من أقوى المتغيرات المساهمة في التمييز بين الاستجابات البشرية واستجابات الذكاء الاصطناعي، في حين أظهرت أحجام التأثير التمييزية لكل من الحساسية للمشكلات (1.464) والأصالة (3.359) أكبر الفروق بين المجموعتين. وتشير النتائج إلى أن استجابات الذكاء الاصطناعي حققت مستويات متقاربة مع استجابات الطلبة الموهوبين في بُعدي الطلاقة والإثراء، إلا أنها جاءت أقل في أبعاد الأصالة، والحساسية للمشكلات، وإعادة التعريف. وتخلص الدراسة إلى أن النموذج المقترح يُعد إطارًا فعالًا وقابلًا للتفسير لتمييز أنماط الإبداع البشري والاصطناعي اعتمادًا على المقاييس النفسية للإبداع.

الكلمات المفتاحية :الطلبة الموهوبون؛ التعلّم التجميعي الهجين؛ التعلّم الآلي؛ قياس الإبداع؛ أهمية الخصائص؛ التعليم العراقي؛ خوارزمية التعزيز التدرّجي المُحسَّنة.