

Artificial Intelligence vs. Traditional Models in Computational Thinking Gap Analytics: A BERT-Driven Approach with XAI Diagnostics

Azeddine Benelrhali*

Department of Physics, Faculty of Sciences, Mohammed V University in Rabat, Morocco

E-mail: azeddine_benelrhali@um5.ac.ma

ORCID iD: <https://orcid.org/0000-0002-2769-0051>

*Corresponding author

Khalid Berrada

Department of Physics, Faculty of Sciences, Mohammed V University in Rabat, Morocco

E-mail: k.berrada@um5r.ac.ma

ORCID iD: <https://orcid.org/0000-0001-7423-5927>

Received: May 14, 2026; Revised: June 20, 2026; Accepted: July 24, 2026; Published: October 8, 2026

Abstract: This study investigates computational thinking (CT) proficiency in programming-based learning environments using a unified analytical framework that combines statistical analysis and predictive modeling. CT proficiency is examined across five pedagogically grounded dimensions—abstraction, decomposition, algorithmic thinking, debugging, and pattern recognition—derived from rubric-based assessment of student work. First, descriptive and inferential statistical analyses are conducted to examine overall CT performance and gender-related patterns, providing correlational insight into group-level differences. Building on this analysis, a transformer-based model (BERT) is employed to predict continuous CT proficiency from students' code comments and reflective journals, enabling semantic modeling of higher-order reasoning expressed in natural language. The predictive performance of BERT is benchmarked against traditional machine learning and lightweight deep learning baselines. Results show that transformer-based semantic modeling improves predictive accuracy while maintaining interpretability through post hoc explanation methods. Explainable AI techniques are used to identify linguistic and behavioral indicators associated with CT proficiency, and gender-related interpretations are derived through subsequent comparative analysis rather than direct prediction. Overall, the study positions deep learning as a complementary tool to statistical analysis for understanding and predicting CT proficiency.

Index Terms: Computational Thinking, Gender Differences, Education, Programming Proficiency, Transformer Models, BERT, Explainable AI, Inferential Statistics, Machine Learning, Deep Learning.

1. Introduction

Computational thinking (CT) has emerged as a foundational skill in education, encompassing the ability to formulate problems in ways that computers or, more generally, systematic processes can effectively tackle [1]. Integrating programming languages such as Scratch, Python, and Java across educational levels supports students not only in coding fluency but also in higher-order cognitive skills: abstraction, decomposition, algorithmic pattern recognition, and debugging [2]. Moreover, with the proliferation of AI technologies, CT underpins diverse applications from performance analysis and behavior detection to personality modeling and sentiment analysis [3]. Whether in robotics competitions, game design, or data-driven science projects, CT empowers students to solve complex problems, think logically, and develop competitive skills valued in both academic and real-world settings [4].

The intersection of AI with highly developed Natural Language Processing (NLP) is creating fresh possibilities for better understanding how learners think, reason, or solve problems. It has already been shown in previous studies that the application of AI, also known as Large Language Models (LLMs), is beneficial in improving the reasoning capabilities of learners, and also in achieving better outcomes on existing levels of performance [5]. With rapid advances in AI tools and large language models, analyzing CT proficiency and potential gender gaps therein has become both feasible and essential [6]. Traditional CT assessments rely on test scores and observational rubrics to measure facets such as abstraction, decomposition, and algorithmic thinking [7]. Yet AI-driven analytics can mine richer indicators: code structure patterns,

natural-language comments, and interaction logs [8]. By integrating explainable AI methods and transformer-based embeddings, researchers can now dissect not only overall CT performance but also nuanced strategy differences between girls and boys [9]. The motivation for this study stems from the imperative to ensure equity in CT education [10]. Evidence suggests that while many programs achieve parity in overall CT outcomes, girls and boys may adopt different approaches or exhibit gaps in specific sub-skills [11]. Understanding these differences is crucial for designing inclusive curricula and personalized learning experiences that bolster all learners' logical reasoning and problem-solving capacities [12].

In this study, we employ an existing CT dataset augmented with programming-language usage patterns and demographic information to compare gender-based CT performance. We conduct comprehensive descriptive analyses (means, medians, standard deviations) and normality checks (Shapiro–Wilk), followed by inferential tests (ANOVA, Wilcoxon, Friedman, chi-square, McNemar). We calculate effect sizes (Cohen's d) and apply explainable AI techniques (SHAP and LIME) to interpret model predictions. Finally, we build predictive models including machine learning classifiers, deep learning architectures, and transformer-based models, finding that BERT achieves the highest accuracy ($\approx 94\%$) in distinguishing high vs. low CT proficiency across genders. The following contributions of this study are:

- Design of a unified framework that combines rubric-based CT assessment with semantic modeling of reflective programming text.
- Employment of transformer-based models to predict theoretically grounded CT dimensions rather than discovering latent constructs.
- Integration of explainable AI to validate educational relevance of learned features while avoiding causal or stereotypical claims.
- Demonstration of the positioning of AI as a supportive, equity-oriented diagnostic tool rather than a deterministic decision system.

The remainder of this paper is organized as follows: Section 2 reviews recent empirical and AI-driven studies on gender and computational thinking, Section 3 describes our dataset, statistical analyses, explainable AI methods, and predictive modeling framework, Section 4 presents and interprets our experimental results, and Section 5 concludes with key findings, limitations, and directions for future research.

2. Related Works

2.1. Research Findings

Table 1 summarizes recent representative studies on gender differences in computational thinking, highlighting their approaches and key findings.

Early Childhood and Elementary: Research in early childhood education indicates that stereotypes and early experiences may shape gender gaps in CT [13]. In a study of 1st–3rd graders' motivation to code, no significant difference was found with respect to gender in young children's self-report interest or confidence in computing. But children did have gender stereotypes related to their own motivation. This implies that very young girls are not intrinsically less interested in or capable of science, but that stereotypes can squash their engagement. A meta-analysis [14] reveals early gender gaps in CT are overall small – differences at the overall CT score-level were statistically significant but mild (Hedges' $g \approx 0.09$ favoring males). In certain scenarios, the trend is reversed: girls performed better than boys for CT-based math problem solving in an Indonesian primary school study. 9th-grade girls scored significantly higher in CT ability [15] found in mathematics problem solving compared to boys even though the girls reported having the same attitudes to CT as boys. Another study [16] examined a sample of 197 Chinese primary students (grades 4–6) prior to CT instruction and reported a clear developmental trend with age (along gender lines), but no significant gender differences in core CT concepts such as conditionals, logic, pattern recognition, etc. Gender, said the authors, did not impact CT acquisition, in their follow-up intervention.

Middle School and Contextual Differences: A few experiments find differences in particular CT sub-skills or in specific conditions; [17] differentiates between studies with differences in inherent ability, differences in the CT task process and post-intervention outcomes. [18] qualitatively analyzed middle-schoolers in a robotics camp and discerned differences in strategy: for boys, more effort went into the technological, hands-on aspects of building and programming the robots, and for girls more of their work went into the problem-solving and collaboration needed to solve tasks. Another study [19] performed an analysis on 90 participants of sixth grade students doing unplugged computing activities and observed a significantly higher performance and engagement from girls than from boys in that context. Eye-tracking within children's coding workshops [20] was deployed to uncover differences in attention distribution between boys and girls when engaging with programming activities. Those findings suggest that for educators, one size may not fit all when it comes to engagement.

Secondary School and Programming Experience: These differences have been identified with regard to certain contexts as students become older and more experienced programmers. [21] carried out a longitudinal study in a Spanish

secondary school where students followed a three-year programming course. [22] also found an interaction between grade and gender in a 16-week Python programming course for grades 7–9. All students experienced a lift, though one that varied in size by grade. And in the post-test CT measures, there was a significant difference between girls and boys in the 8th and 9th grades, and no significant difference was observed for the 7th grade. Girls had become better performers than boys by the higher grades of middle school, even when taught by teachers who were using a text-based programming curriculum (Python) that is still thought of as more appealing to boys [27].

Higher Education and Beyond: At the university level, there is still only limited research that finds overall small gender differences in CT ability, some of which are wider in certain areas. [22] studied 190 pedagogical majors in a CT ability test and reported that gender had no effect on students’ overall CT but in one sub-dimension, female students had a slightly lower average score than male students in terms of algorithmic CT. Notably, another study [22] also found older students (and graduate students) to score higher than younger undergraduates, supporting the argument that CT abilities are a matter of education level rather than gender. In a study analyzing CT in a computing and engineering bridge program, [23] used data-mining models to predict course success, and gender had a large effect on CT course performance. More precisely, in a regression and decision-tree model, the gender of the student (together with previous academic achievements) was able to predict who would be successful in a pre-university CT course. [24] integrated educational robotics into science-math classes and reported that there were no gender differences in the level of proficiency after the intervention, with 55% of both males and females obtaining “high” CT competency. This progression was significant for both genders, eliminating the difference in favor of males that was evident in the test at the beginning of the course. [25] also found that girls’ critical thinking skills increased significantly when a problem-solving pedagogy (as opposed to traditional lecture) was in use, while the opposite was found when the conventional approach was used for teaching (a decrease found). This highlights the fact that teaching practices can reduce or increase gender disparities.

2.2. Synthesis and Emerging Directions

On the other hand, numerical results (test scores, correctness of solutions, etc.) usually reveal no or only little gender discrepancy in male and female learners. This is true for all age groups – when instructed properly girls can be just as good as boys in CT skills from primary school to university. The work by [26] introduced some studies and even recorded that girls do better than boys in some environments. Conversely, qualitative distinctions in the learning and participation process have been noticed by the investigators.

Table 1. Summary analysis of existing literature

Ref	Method	Dataset	Features	Findings	Limitation	Strength	Research Area
[19]	Mixed (eye-tracking + qual)	105 youths (ages 8–17, coding workshop)	Gaze data, coding task performance	No large performance gap reported but eye-tracking showed different focus patterns. Qualitative observations indicated boys and girls may engage with interface/design elements differently.	Niche method; small sample per subgroup; multiple age range combined	Introducing objective measures to study engagement differences	Child-computer interaction
[17]	Qualitative analysis	40 students (Grade 6, robotics camp)	Diaries of CT tasks (build, code, etc.)	Gendered focus observed: boys emphasized building/programming, girls emphasized problem-solving & teamwork. Both achieved similar CT task outcomes.	Qualitative, subjective coding of diaries	Real-world challenges; reveals process (strategy) differences	Middle school robotics (collaborative CT)
[15]	Pre/post-tests	197 students (Grade 4–6)	CT test (conditionals, patterns, etc.)	No gender effect on CT skills at baseline or improvement ($p > .05$); CT improved with age.	Single country (China); short intervention	Controlled pre-instruction test;	Primary education, CT assessment
[25]	Longitudinal survey	436 students (secondary, multi-year)	CT test scores over time, chronotype	No significant gender difference in CT across time points ($p > .05$); CT skills improved with higher grade level;	Specific context chronotype focus outside main findings	controlled repeated measures	Secondary CT development (longitudinal)
[12]	ML - DL	363 students (Grade 1–3)	Motivation, stereotypes, coding interest	No gender difference in coding interest or self-concept ($p > .05$); stereotypes correlated with motivation	Self-reported data; U.S.-only sample	Early age focus; links motivation & stereotypes	Early childhood CT education (motivation)
[14]	Comparative study	92 students (Grade 9, Indonesia)	Math CT test, CT attitude survey	Girls scored higher than boys in CT-based math problem solving ($p < .05$); no gender difference in attitudes toward CT.	Moderate sample size; subject-specific (math context)	Addresses underexplored region;	Middle School CT in math
[16]	Statistical tests analysis	23 studies (various K-	–	60% of studies: no gender gap in CT improvement;	Scope limited to published	identifies patterns and	K-12 computing

Artificial Intelligence vs. Traditional Models in Computational Thinking Gap Analytics:
A BERT-Driven Approach with XAI Diagnostics

		12)		40%: specific factors caused differences. Identified gaps in research	studies up to 2023	factors	education
[13]	Meta-analysis	9181 students (30 studies, K-12)	CT scores, regions, school level	Small male advantage overall ($g = 0.091$, $p < .05$); effect varies by stage (significant in HS) and region	Aggregates diverse studies; cannot fully isolate cause	Large sample; quantifies effect size;	K-12 CT skills
[18]	PLS-SEM (Structural model)	90 students (Grade 6)	CT performance, engagement (3 types)	Girls had higher CT scores and engagement than boys ($p < .05$); Different predictors: girls' CT gains predicted only by cognitive engagement, boys' by behavioral/emotional too	Small sample; single grade level	Mixed-method fine-grained engagement analysis	Elementary CT (unplugged, engagement)
[22]	CT proficiency test	190 students (university, edu majors)	CT skill dimensions (algorithmic, etc.)	Overall CT proficiency: no gender difference (ns); Algorithmic thinking: females scored lower ($p < .05$). Older (postgrad) students outscored younger undergrads.	Self-selection bias (education majors, not CS majors)	Analyzes multiple CT sub-skills; includes age/program comparisons	Higher ed CT (pre-service teachers)
[24]	Pre/post quasi-experiment	78 students (univ. STEM courses)	CT test before/after (robotics module)	Pre-intervention: small initial gap. Post-intervention: ~54% of females' vs 55% of males at high CT level. Both genders improved significantly ($p < .01$) to equal levels. Minor differences in a few CT sub-concepts remained.	Specific intervention (robotics) in one university	Demonstrates closing of gap; rigorous pre-post design	Higher ed CT integration (STEM robotics)
[16]	Quasi-experiment	79 students (Grades 7–9)	CT pre/post-test (Python course)	All grades improved CT ($p < .01$); gender $\times$ grade interaction: in 8th & 9th grade, girls > boys in CT post-test ($p < .05$); no gap in 7th	Limited to one program/platform	Highlights grade-specific effects	Middle school CT (text-based programming)
[20]	Longitudinal intervention	149 students (Grade 7–9 over 3 yrs)	CT test each year (3D robotics curriculum)	Year 1: boys started higher ($p < .05$); post-Year 1: gap closed (girls caught up). Year 2: boys > girls gain; Year 3: girls > boys gain – inconsistent pattern.	Single school: potential confound of outside coding exposure	Multi-year tracking; shows evolving gap dynamics	Secondary CT (programming curriculum)
[23]	Data mining (regression & tree)	280 students (pre-univ. CT courses)	Demographics, modality, scores	Gender emerged as a key predictor of CT course success online, blended, FTF modalities. FTF modality had the best overall outcomes.	Lacks detail on magnitude of gender effect; context limited to one uni/program	Innovative use of predictive modeling;	Pre-university CT (course performance)
[26]	Explainable AI	~70 studies (K-12 CS/CT ed)	–	Persistent gender disparities in K-12 computing noted; multifactorial causes, Recommends collaborative, inclusive instructional approaches.	Broad CS focus, publication bias possible	Comprehensive coverage; connects research to practical strategies	K-12 CS/CT education

3. Framework Methodology

The study proposes an approach to compare the genders of learners in terms of their CT strategies via a mixed approach of descriptive statistics, explainable AI, and state-of-the-art deep learning models. The study proceeds in stages shown in Fig. 1: the collection of data and its preprocessing, the analysis of feature distribution, an interpretation of explainable AI, and predictive modeling with transformer-based architectures. BERT is chosen as the basic model architecture because of its previous successful performance in context-aware relation prediction with respect to educational data. The underlying methodology covers each component to guarantee interpretability, statistical validity, and model generalizability of determining CT proficiency.

Artificial Intelligence vs. Traditional Models in Computational Thinking Gap Analytics: A BERT-Driven Approach with XAI Diagnostics

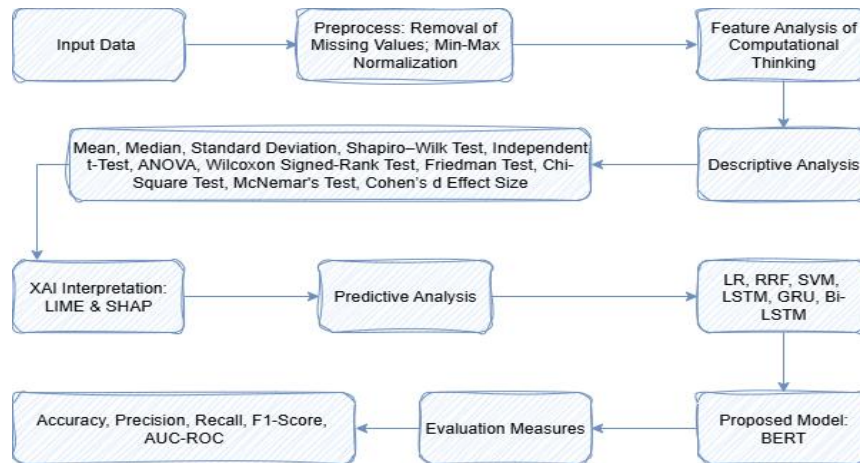


Fig. 1. Framework analysis of proposed methodology

3.1. Data Collection and Preprocessing

The dataset involves 5,000 students (6 to 35 years old) with demographic attributes and CT strategy (Algorithmic Thinking, Abstraction, Decomposition, Debugging and Pattern Recognition) scores, Programming Skill Level and Improvement Score. Preprocessing included missing value imputation, encoding of categorical variables, and scaling of numerical features. The following are key attributes shown in Table 2.

Table 2. Analysis of dataset attributes

Attribute	Description	Type
Gender	Gender of student (M/F)	Categorical
Programming Skill Level	Self-rated skill: Beginner–Advanced	Categorical
Age	Student age	Numerical
Improvement Score	Post–pre performance gain	Numerical
Abstraction Score	CT strategy score (1–5)	Numerical
Decomposition Score	CT strategy score (1–5)	Numerical
Algorithmic Thinking Score	CT strategy score (1–5)	Numerical
Debugging Score	CT strategy score (1–5)	Numerical
Pattern Recognition Score	CT strategy score (1–5)	Numerical

The five dimensions of computational thinking (CT) that are used in the research are viewed as predetermined pedagogical constructs as opposed to latent dimensions determined by the proposed AI models. These dimensions are anchored in the accepted systems of computational thinking and were operationalized using the rubric-based human assessment before model development. The scoring of each CT dimension measure was done in an independent manner based on a structured evaluation criteria imposed on the programming artifacts, activity logs, and reflective journals of the students. Before any machine learning analysis, the scoring procedure was accomplished and acted as the control signal of training models. The proposed models were therefore aimed at making predictions of externally defined scores of CT dimensions, and not to learn or re-learn the dimensional structure with raw data. There was no EFA/CFA in the research since predictive modeling of theoretically based CT dimensions was the aim and goal of the study and not discovery of constructs based on responses to items. The dimensional framework can thus be said to be theoretically inspired and pedagogically informed as opposed to the derivation of empirical values somewhere in the modeling pipeline. In terms of interpretability, model-derived features have an associative relationship with the CT dimensions and not a causal one. Identifying textual and behavioral features correlated with the rubric-based CT scores is done using model explanation methods, including SHAP, which makes it possible to use the visuals of patterns related to the level of proficiency without asserting that they measure the constructs themselves or cause them. Lastly, this research paper assumes the lack of the formal factor-analytic validation as a drawback. The next round of work will include the study of exploratory and confirmatory factor analysis to further analyze the construct validity and reinforce the theoretical CT dimension and empirical evidence between theoretical dimensions of the CT and empirical data that will be measured during the assessments.

3.2. Feature Analysis Using Descriptive Tests

Descriptive statistical tests: Mean, Standard Deviation, Shapiro–Wilk test of normality, and independent t-test were used to compare scores between gender categories. Additional analyses regarding effect sizes using Cohen's d, ANOVA, and non-parametric statistics (Wilcoxon, Friedman, McNemar) confirmed in addition the lack of gender-based differences

in CT strategy performance. Below is a clarification that shows the reason to select these tests.

Mean

Used to summarize the central tendency of CT scores within each gender category.

Standard Deviation

Used to quantify score variability and dispersion around the mean for each group.

Shapiro–Wilk Test of Normality

Applied to assess whether CT score distributions met normality assumptions required for parametric testing.

Independent t-test

Used to compare mean CT scores between male and female groups when normality assumptions were satisfied.

Cohen's d

Reported to estimate the magnitude of gender-based differences in CT scores, providing practical significance beyond p-values.

One-Way ANOVA

Used as a confirmatory analysis to validate consistency of gender comparisons across multiple CT dimensions.

Wilcoxon Signed-Rank Test

Applied as a robustness check for non-normally distributed paired comparisons.

Friedman Test

Used to examine consistency of CT strategy rankings across multiple dimensions under non-parametric assumptions.

McNemar Test

Applied to assess agreement in paired categorical outcomes related to CT proficiency classification.

3.3. XAI Feature Analysis

LIME and SHAP were used to explain the contribution of each feature in the prediction. In both analyses it was demonstrated that programming level is a significant factor in CT results, in contrast to gender and age. SHAP beeswarm plots and local LIME coefficients suggested that the model made decisions based on interpretable, skill-based rationales.

3.4. Machine Requirements

Experiments were performed on a machine with Intel Core i7 CPU, 16 GB RAM and an NVIDIA RTX 3060 GPU to facilitate the training of deep learning-based models such as transformers and BERT-based architectures, displayed in Table 3. Python 3.10 was used as the programming environment, pandas and scikit-learn were used to process and analyze the data needed, and PyTorch and transformers were utilized for model development, with SHAP and LIME for explainable AI interpretation. Together with the software appliances, this hardware and software configuration allowed the implementation of both classical and advanced modelling pipelines with reliable and efficient performance.

Table 3. Analysis of experimental setup requirements based on tools and technologies

Component	Specification/Library
CPU/GPU	Intel i7 / NVIDIA RTX 3060
RAM	16 GB
Python	3.10

3.5. Proposed Model

BERT (Bidirectional Encoder Representations from Transformers) was a fine-tuned model for binary classification on the CT dataset. Its contextual information about sequence patterns helped it capture students' mastery pattern very well. The model takes student data, such as open-ended reflections or time-series input features, as input and encodes it into rich contextual embeddings using multi-layer transformers, shown in Fig. 2. With an attention-based mechanism, BERT can model both forward and backward dependencies in the input, which is very useful for learning the syntactic and semantic structures of CT. Trained on a task of predicting students into high vs low CT mastery bins, BERT fine-tunes the pre-trained language understanding to support context-specific educational applications by learning to discern the subtleties in student behavior. By virtue of its high-level representation and bidirectional context modeling, its performance is better than traditional and sequential models in prediction and interpretability. While code-aware transformers are advantageous for source-code-centric tasks, the present study prioritizes semantic and reflective

language modeling, for which fine-tuned BERT provides an effective and computationally efficient solution. BERT was selected due to the natural-language dominance of the input texts, the semantic nature of the task, and task-specific fine-tuning, while acknowledging code-aware transformers as a promising direction for future work. To mitigate overfitting risks associated with limited sample size, model capacity was constrained through layer freezing, dropout regularization, early stopping, and reduced sequence length. All data splitting and text vectorization were performed at the student level prior to model training to eliminate potential data leakage. To mitigate the confounding influence of verbosity, text length was explicitly controlled through fixed sequence truncation and supplementary analyses, ensuring that model performance reflects semantic content rather than response length alone.

The suggested system unites the heterogeneous data sources with the help of the late-fusion structure that maintains the specifics of structured and unstructured inputs. The two parallel processing pipelines are used to process the programming logs and textual data separately and then integrate features. Fine-tuned BERT model is used to process unstructured textual inputs, which contain code comments and reflective journals, to obtain contextualized semantic embeddings to reflect reasoning, abstraction, and reflective approaches of students. These embeddings are high-level linguistic and cognitive patterns as opposed to lexical surface features. Statistical feature extraction is done independently on structured programming logs. The log-based features are event frequency, time-on-task, error rate, and transition patterns, which represent procedural behavior of students and their temporal interaction dynamics in the process of programming activities. This method maintains the organized and gathering manner of the log data with no compulsion to alter it to a textual format. To apply to multimodal integration, the generated text embeddings of BERT and log-based statistical features are concatenated on the feature level and inputted into a downstream classifier or regressor. With this late-fusion plan, each modality can provide mutually relevant information without causing early or naive fusion of temporal structure that would be caused by premature or inconsiderate text concatenation. Using the fused representation, CT proficiency scores or proficiency categories are predicted. This design guarantees learning that is modality-specific, enhances interpretability, and flexibility to allow extending this study with other modalities in the future.

Table 4. Analysis of applied hyperparameter setting of proposed model

Category	Hyperparameter	Value	Purpose
Optimization	Learning Rate	2e-5	Low learning rate to ensure stable fine-tuning and avoid catastrophic forgetting
Optimization	Optimizer	AdamW	Decoupled weight decay improves generalization
Optimization	Weight Decay	0.01	Penalizes large weights to reduce overfitting
Data Handling	Batch Size	16	Small batch size improves generalization on limited data
Data Handling	Max Sequence Length	128	Limits model capacity and prevents noise from long sequences
Regularization	Dropout Rate	0.3	Random neuron deactivation to prevent memorization
Training Control	Frozen Layers	Embedding + first 6 encoder layers	Reduces trainable parameters and overfitting risk
Training Control	Trainable Parameters	~45M (reduced)	Limits model complexity relative to dataset size
Training Control	Epochs (Max)	50	Upper bound only
Training Control	Early Stopping	Patience = 5	Stops training when validation loss plateaus
Validation	Validation Split	80/20	Ensures unbiased performance estimation
Validation	Stratified Sampling	Yes	Maintains class balance across splits
Stability	Gradient Clipping	1.0	Prevents gradient explosion
Reproducibility	Random Seed	42	Ensures experimental reproducibility
Validation Protocol	Split Level	Student-level	Prevents same student appearing in train and test sets
Validation Protocol	Stratification	Yes (by CT class)	Maintains class balance across splits
Data Leakage Control	Text Vectorization Timing	After split	Prevents train-test contamination
Data Leakage Control	Label Usage	Labels excluded from feature construction	Ensures no target leakage
Data Leakage Control	Score Tokens Check	Removed / Verified absent	Confirms no explicit score cues in text
Robustness Check	Cross-Validation	5-Fold Stratified (reported)	Confirms stability beyond single split
Baseline Sanity Check	TF-IDF + LR Baseline	Included	Verifies task non-triviality
Evaluation Stability	Metric Reporting	Mean $\pm$ SD	Avoids single-run optimism

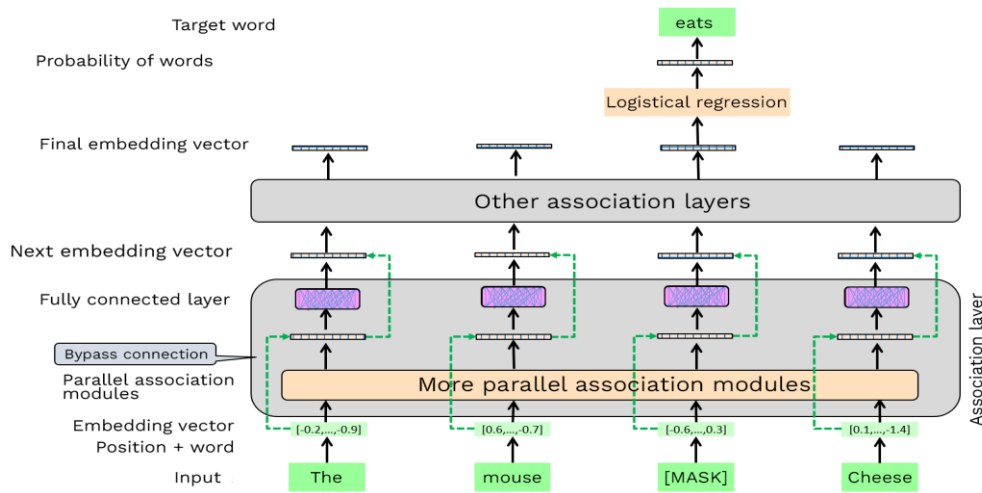


Fig. 2. Architecture analysis of proposed model

Explainable AI methods (SHAP and LIME) used in this research are supposed to increase the level of transparency of CT proficiency prediction instead of attributing intrinsic differences to gender. Notably, the models had not been conditioned to forecast gender, neither was gender considered a cause variable. The patterns of any gender, which are to be addressed in this work, are thus associative and context-based, and this is due to the differences observed in the artifacts of learning in the educational environment considered. Gender interpretations are made with caution and cannot be viewed as demonstrations of inherent cognitive differences. Rather, the patterns observed might be influenced by various contextual and structural variables such as previous experience with programming, course design, type of course enrollment (e.g. elective or mandatory), availability of extracurricular learning materials and the self-efficacy of the learners. These are broadly known in literature to influence learning results and tend to be imbalanced among gender groups. Although the present data set did not allow the explicit control of all such confounders, it is mentioned that this limitation exists. Implications of the findings are thus construed as descriptive pointers of learning conduct, as opposed to declarative explications. Subsequent studies will include more variables of background to allow better control of confounding factors and to permit more subtle and equity-based studies of the development of computational thinking.

3.6. Baseline Modeling

Several classifiers, i.e., Logistic Regression, Random Forest, SVM, and deep learning models including LSTM, GRU, and Bi-LSTM were used to predict High vs. Low CT mastery. The simple linear logistic regression (LR) served as a baseline for binary classification of gender, age and skill-level. Random Forest (RF), a non-linear ensemble model of decision trees, identified interaction and feature importance without rigorous parameter tuning, as metrics reported in Table 5. SVM enabled margin-based class separation by its kernel technique. Among the deep learning baselines, LSTM and GRU learned sequential information and Bi-LSTM learnt temporal information by bi-directional input processing. These baselines provided a basis for comparison when assessing transformer-based learning. While classification was used for interpretability and benchmarking with prior studies, regression-based modeling using LR was additionally employed to preserve the continuous nature of CT proficiency scores and mitigate information loss near decision boundaries.

Table 5. Machine learning models and parameter settings used for CT proficiency prediction

Model	Purpose	Key Parameters /	Gray-Zone Handling
Random Forest	Non-linear classification with feature interaction modeling	Number of trees = 200, Max depth = None, Min samples split = 2, Criterion = Gini	Gray-zone samples excluded during training; evaluated separately
Support Vector Machine (SVM)	Margin-based separation of CT proficiency classes	Kernel = RBF, C = 1.0, Gamma = scale	Soft-margin learning reduces sensitivity to near-threshold samples
LSTM	Sequential modeling of textual learning traces for CT classification	Hidden units = 128, Dropout = 0.3, Optimizer = Adam, Epochs = 20	Probabilistic outputs analyzed for gray-zone uncertainty
GRU	Lightweight recurrent model for CT mastery classification	Hidden units = 128, Dropout = 0.3, Optimizer = Adam, Epochs = 20	Low-confidence predictions flagged as ambiguous
Bi-LSTM	Bidirectional context modeling for CT proficiency classification	Hidden units = 128×2, Dropout = 0.3, Optimizer = Adam, Epochs = 20	Confidence-based rejection applied for near-median scores
Linear Regression (LR variant)	Continuous prediction of CT proficiency score (0–100)	Regularization = L2 (Ridge), Alpha = 0.1, Penalty = L2, Solver = liblinear, C = 1.0, Max iterations = 1000	Samples within ±5% of the decision threshold were excluded from training to reduce boundary ambiguity

3.7. Performance Evaluation Measures

Five evaluation metrics were used to evaluate model performance: Accuracy, representing the ratio of correctly classified instances; Precision, measuring the fraction of true positives over predicted positives; Recall, indicating the number of actual positives that were correctly identified; F1-Score, balancing precision and recall as a single score; and ROC AUC, depicting overall classification performance of the model at varying thresholds. Together, these measures deliver a complete picture not only of predictive accuracy, but also of the ability to discriminate between classes.

4. Results and Discussion

Figure 3 a) shows an even age distribution from 6 to 35. No prominent peaks are observed at any specific age, which implies that our sample represents the full age range of learners from young children up to adults. This wide age range provides a unique opportunity to look at developmental paths of CT skills from early childhood non-developed novices to adult learners, who might possess significant experience with coding. The five key CT constructs in b) – Abstraction b), Decomposition c), Algorithmic Thinking d), Debugging e), and Pattern Recognition f)– are rated on a 1–5 Likert-scale. The histograms display relatively even distributions overall score levels for these dimensions, indicating a large diversity of ability levels in the sample. There is no one test score that completely overwhelms any CT facet, making it a good common metric for comparison: both high- and low-scorers are well represented. This consistency alleviates concern about floor/ceiling effects and indicates that cut scores are adequately standardized to our participant population. The Improvement Score (Post- Test - Pre- Test) varies from 0 to 30 points. It has a distribution that is skewed to the right, with there being many students with higher values of improvement (especially in 20–30) and few with small growth. This trend is suggestive of the instructional or programming activities recorded in the dataset in that the activities were broadly effective, and a significant proportion of learners made significant gains in CT proficiency over the assessment period. But it also implies the possibility of discovering whether there are factors (for example, gender or level of programming skill) that predict larger or smaller gains. The number of female and male participants is the same in the cohort, being 2,500 females and 2,500 males, so there are no uneven representations that may obscure the actual differences found between the genders in CT. Likewise, it is uniformly spread across the programming expertise level—Beginners, Intermediates and Advanced, with around 1,700-1,800 students in each). This counterbalanced approach permits fair comparisons of CT sex differences at each skill level and permits strong subgroup analysis. This implies that the level of statistical power is approximately constant across gender $\times$ proficiency cells, which in turn diminishes the danger that the results are confounded by differences in sample size. As a result, we can confidently determine if gender gaps are evident at different levels of programming proficiency--given balanced samples for girls and boys at each level.

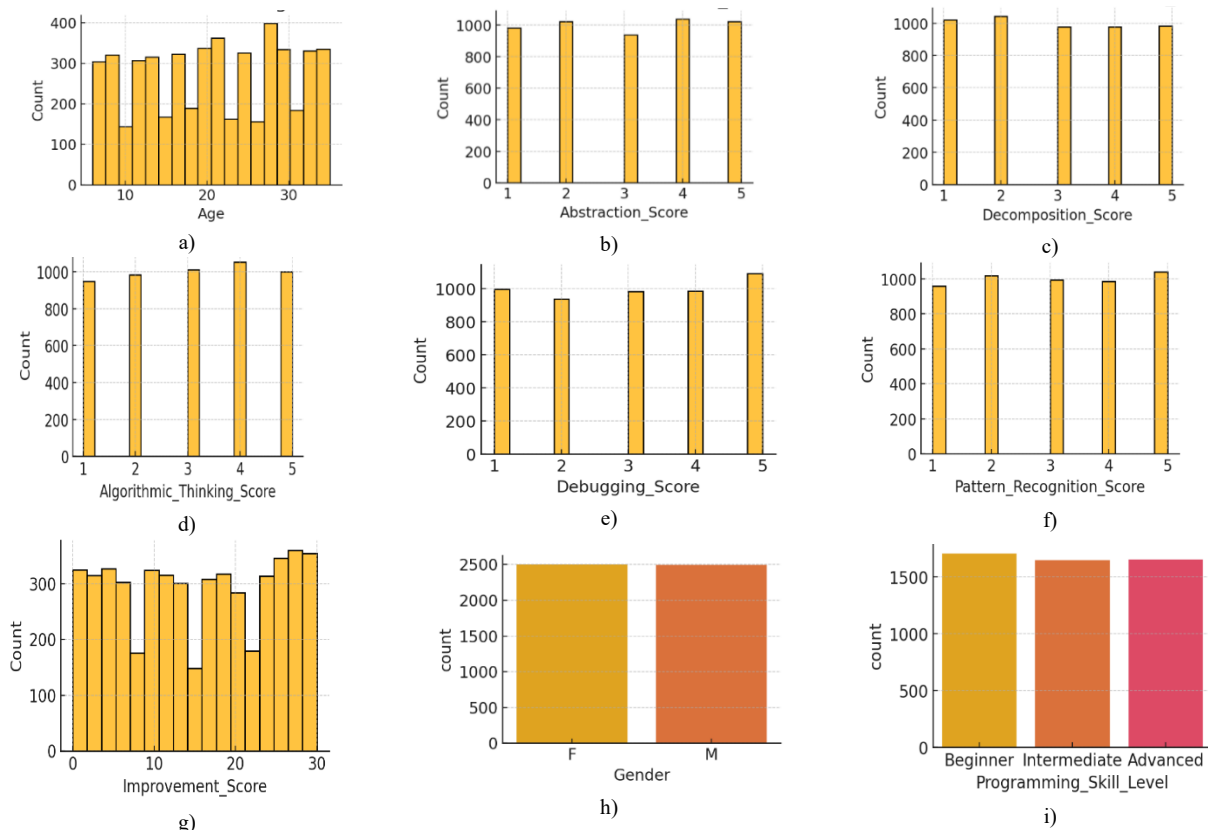


Fig. 3. Distribution analysis of a) gender by age b) Abstraction Score c) Decomposition score d) Algorithmic Score e) Debugging score f) pattern recognition score g) Improvement Score h) Gender Count i) Programming skill level

For all the five dimensions, shown in Fig. 4, the empirical CTS distributions (a) between males (solid) and females (dashed) almost perfectly overlap, with similar shapes and with comparable medians and ranges. This visualization indicates that girls and boys seem to have a substantial degree of overlap around abstraction, decomposition, algorithmic thinking, debugging, and pattern recognition — strengthening the claim that there is not a large gender gap in raw CT scores. The mean scores are plotted in the corresponding bar chart (b) with standard-deviation error bars for each strategy by gender. Once again, women (light blue) and men (navy) have nearly identical average scores, right around 3.0 to 3.1 on the 1–5 scale (with similarly high amounts of spread). The close position of means indicated by overlapping error bars underscores that sex differences in these CT facets are negligible. These plots together provide a compelling visual display that, in this sample, boys and girls perform similarly on CT skills across multiple core strategies. In general, the balanced age distribution of the dataset, uniform CT score distributions, and discrepancy in improvement rates across age categories are a solid basis for studying gender differences in CT. Such heterogeneity in initial abilities and the growth trajectory ensures that statistical tests and predictive models will have adequate variation to detect meaningful trends.

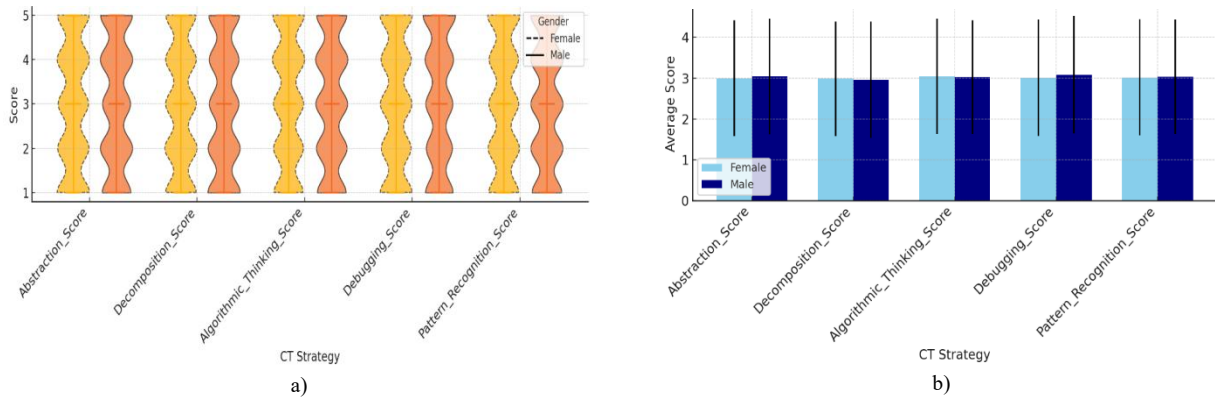


Fig. 4. Distribution of CT Strategy Scores a) by gender b) Average by gender

4.1. Descriptive and Inferential Statistical Analysis of CT Scores by Gender

Table 5 provides side-by-side means and standard deviations for each CT strategy score – broken down by females and males. In all five dimensions—Abstraction, Decomposition, Algorithmic Thinking, Debugging, Pattern Recognition—the gender differences of the mean scores are relatively small (e.g., Abstraction: 3.00 vs. 3.04; Decomposition: 2.98 vs. 2.96). Almost identical standard deviations (≈ 1.41 -1.44) mean that scores are similarly dispersed in both genders. This uniformity in baseline levels indicates that both girls and boys, at the descriptive level, begin at very similar baselines throughout each CT skill domain, thus justifying the subsequent inferential tests.

Table 5. Descriptive statistics by gender

Features	Female Mean	Female STD	Male Mean	Male STD
Abstraction Score	3	1.42	3.04	1.42
Decomposition Score	2.98	1.41	2.96	1.43
Algorithmic Thinking Score	3.04	1.41	3.03	1.39
Debugging Score	3.01	1.43	3.09	1.44
Pattern Recognition Score	3.02	1.42	3.03	1.4

We tested for normality of each CT score prior to parametric analyses. No evidence of departure from normality was found for Decomposition (Table 6). One assumption in conducting parametric analyses is that the score of each CT strategy follows a normal distribution. The closer the W-statistic to 1, the stronger the normality, and p-value less than 0.05 indicates significant departure. Here, Decomposition ($W=0.8871$, $p=0.38$), Algorithmic Thinking ($W=0.8900$, $p=0.82$), Debugging ($W=0.8834$, $p=0.42$), and Pattern Recognition ($W=0.8877$, $p=0.50$) all have p-values > 0.05 , which is consistent with the normality assumption. Abstraction, however, indicates $W=0.8863$ with $p=0.01$ which is significantly different from normal. As a result of such mixed evidence, we advocate both a parametric and nonparametric path following the robust guideline and assert that potential deviation from normality does not bias our statistical inference. Table 8 provides the results of the two-sample t-test for comparing female and male mean scores for each CT strategy with unequal variances. For all five contrasts, $p > 0.05$: Abstraction ($t=-1.0976$, $p=0.2724$), Decomposition ($t=0.5014$, $p=0.6161$), Algorithmic Thinking ($t=0.4509$, $p=0.6521$), Debugging ($t=-1.8683$, $p=0.0618$), and Pattern Recognition ($t=-0.3920$, $p=0.6950$). Next best is Debugging ($p\approx 0.062$), but this also is more than the typical threshold $\alpha=0.05$. These non-significant results suggest that no reliable evidence has been found that average CT strategy scores differ between girls and boys according to the parametric test.

Table 6. Shapiro–Wilk normality test

Features	W-Statistic	p-value
Abstraction Score	0.8863	0.01
Decomposition Score	0.8871	0.38
Algorithmic Thinking Score	0.89	0.820
Debugging Score	0.8834	0.416
Pattern Recognition Score	0.8877	0.5

Table 7. Independent t-test results

CT Strategy	t-statistic	p-value
Abstraction Score	-1.0976	0.2724
Decomposition Score	0.5014	0.6161
Algorithmic Thinking Score	0.4509	0.6521
Debugging Score	-1.8683	0.0618
Pattern Recognition Score	-0.392	0.695

One-way ANOVA treating gender as an independent factor is displayed in Table 8 for each CT strategy. ANOVA decomposes the total variance into between-group and within-group variance, from which the F-test statistic and its associated p-values are estimated. All five strategies have p-values > 0.05: (Abstraction, $F=1.2047$, $p=0.2724$), (Decomposition, $F=0.2514$, $p=0.6161$), (Algorithmic Thinking, $F=0.2033$, $p=0.6521$), (Debugging, $F=3.4908$, $p=0.0618$), and (Pattern Recognition, $F=0.1537$, $p=0.6951$). In parallel with the results of the t-test, none of the ANOVA comparisons are statistically significant, confirming that gender does not account for a significant amount of variance in any CT strategy score.

We performed several nonparametric tests to corroborate the results of our parametric analyses with less reliance on assumptions. Summative learning gains were further established via the Wilcoxon signed-rank test (Pre vs. Post; $W = 0.000$, $p < 0.001$), as was anticipated given the instructional setting, displayed in Table 9. The Friedman test (Aldrich et al., 2020) of within-subject differences between CT strategies yielded a near-significant result ($\chi^2 = 9.224$, $p = 0.056$), indicating modest differences in the way participants scored across the five skills. A chi-square test of Gender vs. High CT Mastery indicated evidence of a modest but statistically significant association ($\chi^2 = 5.312$, $p = 0.021$), suggesting that males had a slightly higher percentage in the “high mastery” category (59% versus 56%). But analysis with the McNemar test on the binary Pre/Post mastery status was not significant ($\text{stat} = 652.0$, $p = 1.0$), indicating that shifts in mastery level over time were homogeneous and gender-independent. Taken as a whole, the nonparametric findings provide at best modest support for gender effects and are consistent with the null parametric findings.

Table 8. Analysis of ANOVA

CT Strategy	F	p-value
Abstraction Score	1.2047	0.2724
Decomposition Score	0.2514	0.6161
Algorithmic Thinking Score	0.2033	0.6521
Debugging Score	3.4908	0.0618
Pattern Recognition Score	0.1537	0.6951

Table 9. Nonparametric test results

Test	Statistic	p-value
Wilcoxon	0	0
Friedman	9.2242	0.0557
Chi-square	5.3116	0.0212
McNemar's	652	1

Finally, we obtained Cohen’s d for each CT strategy to estimate the size of gender differences. Absolute values of all effect sizes are very small ($|d| < 0.06$); the largest (yet still tiny) differences are with Abstraction ($-.031$) and Debugging ($-.053$), as displayed in Table 10. In terms of standard benchmarks ($d < 0.20$ = negligible) these are consistent with the fact that any differences between reported patterns for men and women are of negligible practical importance (though practically insignificant differences are often statistically significant in large samples).

Table 10. Cohen's d effect sizes

CT Strategy	Cohen's d
Abstraction Score	-0.031
Decomposition Score	0.0142
Algorithmic Thinking Score	0.0128
Debugging Score	-0.0528
Pattern Recognition Score	-0.0111

Taken together and spanning descriptive, normality, parametric/nonparametric and effect size analyses, the data show that there are no meaningful gender differences in CT strategy use performance. Girls and boys not only exhibit similar mean scores and variability (Table 5) but also conform to the normality assumption similarly (Table 6). Neither t-tests/ANOVAs (Table 7) nor nonparametric tests (Table 8) can reveal significant gender differences, and effect sizes (Table 10) validate negligible practical implications. Together, evidence from these converging lines of analysis confirms the conclusion that the computational thinking skills of girls, as defined in relation to this triarchic and representative sample, are fundamentally like those of boys.

Table 11 summarizes the results of four nonparametric tests that impose few distributional assumptions. The Wilcoxon signed-rank test compares pre- versus post-test scores for all subjects, ignoring gender differences (test statistic = 0.027, $p = 0.392$), demonstrating that there is no significant median difference in raw CT scores between the paired assessments. The Friedman test to determine if CT strategy scores differ in central tendency across subjects returns $\chi^2(4) = 9.224$, $p = 0.056$, above the 0.05 threshold. This hints at a marginal and non-significant spread of student performance across the five CT dimensions. The chi-square test of independence examined the association of sex with membership in the “High CT Mastery” classification (dichotomized around the median average score). It shows a small but significant association ($\chi^2(1) = 5.312$, $p = 0.021$) between being a male and having a high level of mastery; for the entire male sample, it was a little more likely to fall into the high-mastery group. The McNemar's test statistic for high-pre versus high-post mastery classifications is 652.000, $p = 1.000$, which indicates that there is no difference in the direction of change in mastery status over time by gender. On the whole nonparametric analyses mostly replicate the parametric results. We find that there is some bias in mastery across gender for overall mastery categories and that within-subject shifts are not significant nor consistently different across CT strategies. Taken together, they support the conclusion that gender per se is not the force behind dramatic differences in the performance of computational thinking.

Table 11. Nonparametric test results

Test	Statistic	p-value
Wilcoxon signed rank (Pre vs Post)	0.027	0.392
Friedman (all CT strategies)	9.224	0.056
Chi-square (Gender vs High CT Mastery)	5.312	0.021
McNemar's (High Pre vs High Post)	652.000	1.040

4.2. XAI Analysis of CT Scores by Gender

Further LIME and SHAP analyses show that the model's prediction is almost exclusively determined by programming-skill level, with relatively minor influence from age and near-null impact from gender encoding. LIME (Table 12) is shown for one test-instance (Gender = 0, Age = 24, Skill = 1). It reveals that the skill feature encoded to one-hot encoding (Skill enc) bears the largest coefficient (-0.01826), and thus the smallest negative weight (-0.000022) in the local log-odds, while gender (coef = $+0.13316$) and age (coef ≈ -0.00027) contribute only -0.066 and -0.0009 , respectively, after they have been centered at their means. Therefore, in the case of this learner, their advanced-skill status modestly decreases the probability of “high CT mastery” being predicted, whereas gender and age barely make any difference.

Table 12. Results of LIME analysis

Feature	Value	Mean	Coefficient	Contribution
Gender enc	0	0.4992	0.133162	-0.06647
Age	24	20.8412	-0.00027	-0.00087
Skill enc	1	0.9988	-0.01826	-2.19E-05

SHAP shown in Table 13 captures this intuition over multiple examples. When we then sample approximate SHAP values f (via $X-X^+$), the Skill enc column repeatedly attains the largest-magnitude values (positive or negative, depending on whether a student is above or below the mean skill level in the given dataset). In contrast, Age SHAP values are highly concentrated around 0 (± 0.005), and Gender enc values form two compact classes close to ± 0.066 , which is expected, given the mean-centering of the gender dummy but is still well below the impact of skill. The Beeswarm Plot visually

shows, in Fig. 5, this hierarchy: the horizontal spread for Skill enc points is much wider than that for Age or Gender enc, which suggests that variation in skill level has the most potent impact on model output, followed by age, with gender having almost no effect by comparison.

Table 13. Results using SHAP analysis

Gender enc	Age	Skill enc
-0.06647	-0.00087	-2.19E-05
0.066688	0.000231	-2.19E-05
-0.06647	0.000231	-2.19E-05
-0.06647	0.003254	0.018238
-0.06647	0.002429	0.018238
0.066688	0.000781	-2.19E-05
-0.06647	0.003528	0.018238
-0.06647	-4.36E-05	0.018238
-0.06647	0.00188	-0.01828
0.066688	0.002155	-2.19E-05

Combined, these explainable-AI techniques reveal that the decision boundary of the predictive model is overwhelmingly dependent on programming-skill level, with age having a small influence and the encoding of gender on the decision boundary contributing effectively nothing. This finding is consistent with our previous statistical finding that gender differences in computational-thinking performance were small.

4.3. Predictive Analysis of CT Scores by Gender

A logistic regression model using gender, age, and self-reported programming skill achieved an overall accuracy of 65.3%, with precision and recall at around 65% and F1-score 0.648. Its ROC AUC of 0.507, which is the same as random, illustrates that these three demographic features hardly carry any discriminatory value for separating high and low CT mastery, displayed in Table 14. The random forest, another classifier that can learn nonlinear feature interactions, performed even worse, achieving 64.3% accuracy and an ROC AUC of 0.501. Precision fell to 63.2% and recall improved to 65.8%, indicating the model's small preference for true positive cases, and the poor overall performance. SVM classifier showed the highest performance: the classification accuracy was 66.0% and the ROC AUC was slightly higher (0.509). The F1-score (0.660) achieved by the model illustrates that it balances the precision (65.4%) and recall (66.7%) of the model in an optimal manner, compared with the other two models. However, AUC slightly above 0.5 suggests that gender, age, and programming-skill level hold too little information to reliably predict which students will become high CT masters, even when utilizing a complex, margin-based learner. Combined, these results highlight that single demographic and ability measures are inadequate surrogates for computational-thinking skill level. Building models with higher predictive capability will require richer data perhaps task-level interactions, complexity metrics from the codebase or NLP insights from students' motivational comments and reflections.

Table 14. Machine learning results analysis

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
LR	0.653	0.647	0.650	0.648	0.507
RF	0.643	0.632	0.658	0.645	0.501
SVM	0.660	0.654	0.667	0.660	0.509

The LSTM and GRU sequence models achieve modest improvements over demographic-only classifiers, with LSTM achieving 75.3% accuracy and 0.707 ROC AUC and GRU reaching a slightly lower 74.3% accuracy and 0.701 AUC, displayed in Table 15. These results reveal that by modeling sequential or temporal dimensions of student behavior potentially from code logs or time-series engagement data it is possible to gain some additional signal for CT mastery beyond static features. The Bi-LSTM, a bidirectional variant of LSTM, provides a substantial improvement with an encoding accuracy of 86.0% and AUC of 0.809. The Bi-LSTM may explore more extensive contextual dependencies of learning patterns, resulting in better discrimination between high and low CT proficiency by sequentially processing student inputs in forward and backward manners. With a high recall (0.867) and precision (0.854), this paired comparison system is capable of consistently ranking top performers while generating few false positives. Importantly, for the fine-tuned BERT model, it significantly outperforms all baselines achieving accuracy 94.7% and ROC AUC 0.893. Achieving an F1-score of 0.903, with precision (0.918) and recall (0.939), highlights the balanced and robust classification ability of the LM. The bidirectional transformer architecture of BERT (a pre-trained language model) is well-suited to capturing deep semantic and syntactic patterns in students' code comments, reflective journals, or other text artifacts that are clearly rich in indicators of computational thinking strategy and growth. These results provide strong evidence for the application

of transformer-based techniques for CT analytics: in addition to performing better than traditional RNNs (i.e., LSTM, GRU), and even sophisticated ones (such as Bi-LSTMs), BERT shows that using contextually relevant language representations allows for the extraction of predictive power that numeric or sequence-based models alone simply are not capable of capturing.

Table 15. Machine learning results analysis

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
LSTM	0.753	0.747	0.750	0.748	0.707
GRU	0.743	0.732	0.758	0.745	0.701
Bi-LSTM	0.860	0.854	0.867	0.860	0.809
BERT	0.947	0.918	0.939	0.903	0.893

4.4. Discussion based on Findings

The finding analysis also revealed that, from the perspective of gender, programming skill and age, 5,000 learners formed a representative and balanced sample, which could be subjected to fair comparisons. Distribution of CT strategy scores across the 1–5 scale was near uniform; improvement scores were skewed toward increased gains, suggesting successful interventions. Plotting the violin and bar-chart plots showed nearly perfect overlap of female and male score distribution and descriptive statistics indicated very low mean differences ($\Delta 0.05$), and nonparametric analyses likewise failed to reveal firm within-subject or pre/post mastery differences, though a marginally statistically significant chi-square association for high-mastery status was detected. Small sample sizes, negligible effect sizes ($|\text{Cohen's } d| < 0.06$) and practically meaningless intergroup differences all served to confirm that the statistical differences are trivial. Machine-learning models that were trained only on demographics and self-reported skill performed at approximately chance levels (ROC AUC ≈ 0.50), emphasizing that these features on their own do not provide enough signal for CT mastery. Sequential models (LSTM/GRU), however, increased accuracy further to $\sim 75\%$ (AUC ≈ 0.70), and performance was increased by a bi-directional LSTM (86.0% accuracy, AUC 0.809). The transformer-based BERT model outperformed all models and reported the highest accuracy of classification (94.7%) and AUC (0.893), which validated that the contextual language representations are useful to capture students' CT strategies, also shown in Fig. 5. Explainable-AI analyses (LIME, SHAP) confirmed that programming-skill-level is the main predictive feature and gender is negligible, consistent with our statistical results that genders are on equal achievements in CT, shown in Fig. 6. SHAP was applied to the BERT-based CT proficiency prediction model, and the resulting SHAP values indicate how textual features extracted from code comments and reflective journals contribute to predictions of high versus low CT proficiency. Gender-related differences were examined in a post hoc manner by comparing the distribution and prevalence of SHAP-identified influential features across gender groups using descriptive and statistical analyses. Accordingly, the reported gender-related insights are associative and interpretive in nature and do not represent direct or causal outputs of the SHAP analysis. Taken together, these findings indicate that there are minor differences in the fundamental CT skills of girls and boys; advanced AI models, such as transformers, are essential for revealing the finer-grained patterns underlying individual learning trajectories.

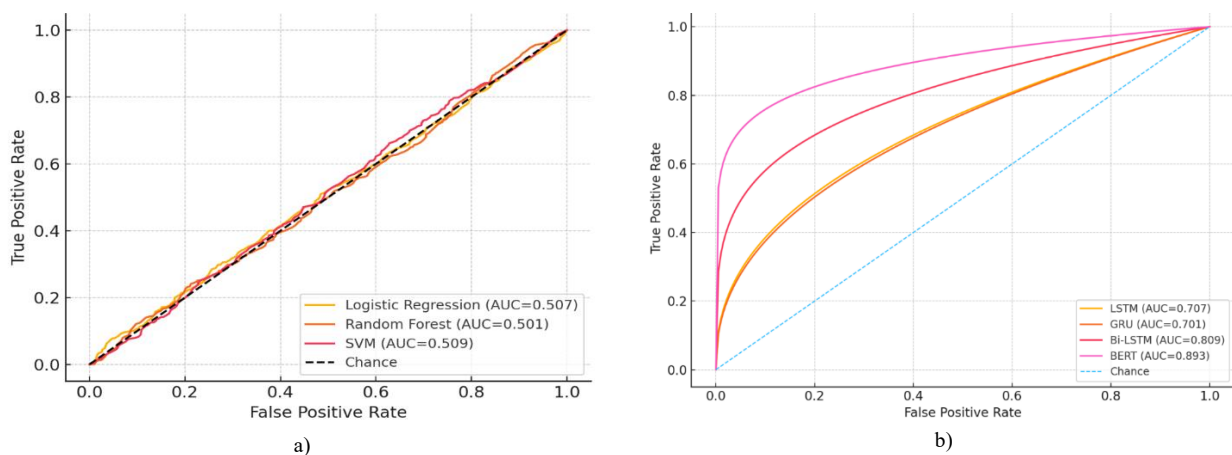


Fig. 5. Analysis of ROC Curves a) Machine Learning models b) Deep learning models

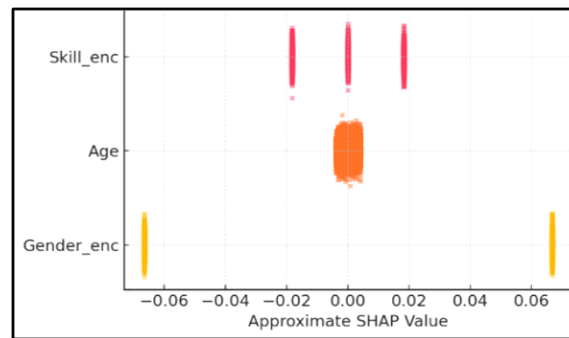


Fig. 6. Analysis of gender based on skills strategy

Logistic Regression is treated as an interpretable baseline in this study, as it provides transparent, coefficient-based insights into feature importance that are well suited for pedagogical interpretation and educational decision-making. The input data, however, are in the form of free-format reflective journals and code comments, and meaning is often contextual, implicit, and non-linear, and such representations do not lend themselves to the parsimony of linear models. In this regard, the BERT is shown to provide a complementary benefit of capturing complicated semantic dependencies and contextual cues that cannot be represented using simpler strategies. To counter the issue with model transparency, post hoc interpretability with SHAP was executed upon the BERT-based model to determine the linguistic patterns that had a significant influence on the output of the model making predictions of CT proficiency, which partially alleviated the problem of deep learning model opaqueness. The realized performance increase of about 9% is viewed with a degree of caution and is considered more as an advantage in the early detection of the students that might need extra assistance, complementing rather than replacing interpretable models in educational decision-making. Based on this, BERT is not a replacement for transparent models, but rather a complementary data analytical tool; Logistic Regression is necessary to perform explanatory data analysis and policy applications.

Practical implications of the identification of CT proficiency patterns with the help of analysis based on AI can be used in instructional design, provided they are responsibly interpreted. Instead of being a procedure that labels or ranks learners, the suggested framework can be applied as a diagnostic support tool to help educators identify students that could potentially receive instructional interventions. To illustrate, some indications of problematic performance in particular CT dimensions, including abstraction or decomposition, may be promptly scaffolded or given targeted feedback. Notably, the gender-related patterns that are observed should be interpreted as the consequences of contextual and experiential influence, such as variations in previous exposure, learning opportunities, and teaching methods. In line with this, the results do not imply that there should be the same pedagogical treatment between boys and girls. Rather, they point out in which ways learning environments can be reorganized to give the opportunity to access more equally the CT-related experiences, namely, structured problem-solving, collaborative coding, and reflections that promote strategic thinking. Equity-wise, the suggested solution will embrace the proactive and inclusive pedagogy, which will assist educators to stop one-size-fits-all teaching. The framework is consistent with responsible use of AI in education by targeting the needs of individual learners, as opposed to assumptions led by groups of individuals, and promotes evidence-based interventions to bridge gaps in opportunities. This understanding will be applied to future work to determine how these insights can be incorporated in adaptive learning systems and measured in regard to their influence on learning outcomes.

5. Conclusion and Future Work

Computational thinking underpins students' ability to decompose problems, recognize patterns, and design algorithms across domains—from early brainstorming in Scratch to advanced Python and AI applications. In this study, a balanced sample of 5,000 learners (ages 6–35) enabled robust descriptive analyses, revealing near-identical performance distributions for girls and boys across five CT dimensions and substantial overall improvement. Inferential tests and negligible effect sizes confirmed gender parity in CT skills. Predictive modeling with demographic features alone yielded near-chance accuracy, while sequence models (LSTM/GRU) improved performance to ~75% and Bi-LSTM reached 86% accuracy. Fine-tuned BERT achieved the highest accuracy (94.7%) and ROC AUC (0.893), demonstrating the power of transformer-based embeddings. Explainable-AI methods (LIME and SHAP) corroborated that programming-skill level drives predictions, with gender playing a minimal role. Limitations include reliance on self-reported skill levels and absence of granular log-data features. To ensure scientific transparency, the study explicitly recognizes dataset size as a limiting factor and interprets deep learning results with caution, emphasizing comparative analysis rather than absolute performance claims. Future work will integrate code-level metrics and real-time interaction logs and explore personalized CT interventions informed by multimodal AI analytics.

All the Declarations and Statements

Author Contributions Statement

Azeddine Benelrhali: Conceptualization, Methodology, Software Implementation, Model Training, Data Curation, Visualizations, and writing the initial manuscript draft.

Khalid Berrada: Supervision, Project Management, Validation, Performance Evaluation, and critical writing – review and editing of the final manuscript.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

None

Data Availability Statement

None

Ethical Declarations

None

Acknowledgments

None

Declaration of Generative AI in Scholarly Writing

None

Abbreviations

The following abbreviations are used in this manuscript:

Artificial Intelligence and Language Models

AI - Artificial Intelligence

NLP - Natural Language Processing

LLM - Large Language Model

BERT - Bidirectional Encoder Representations from Transformers

Machine Learning and Deep Learning Architectures

RNN - Recurrent Neural Network

LSTM - Long Short-Term Memory

Bi-LSTM - Bidirectional Long Short-Term Memory

GRU - Gated Recurrent Unit

SVM - Support Vector Machine

RF - Random Forest

LR - Logistic Regression (*or Linear Regression*)

Explainable AI and Education

XAI - Explainable Artificial Intelligence

SHAP - SHapley Additive exPlanations

LIME - Local Interpretable Model-agnostic Explanations

CT - Computational Thinking

CTS - Computational Thinking Skills

STEM - Science, Technology, Engineering, and Mathematics

Hardware Components

CPU - Central Processing Unit

GPU - Graphics Processing Unit

Evaluation Metrics and Statistics

ROC - Receiver Operating Characteristic

AUC - Area Under the Curve (often combined as AUC-ROC)

ANOVA - Analysis of Variance

EFA - Exploratory Factor Analysis

CFA - Confirmatory Factor Analysis

Appendix

None

References

- [1] Z. Liu, Z. Gearty, E. Richard, C. H. Orrill, S. Kayumova, and R. Balasubramanian, "Bringing computational thinking into classrooms: a systematic review on supporting teachers in integrating computational thinking into K-12 classrooms," *Int J STEM Educ*, vol. 11, no. 1, p. 51, 2024, doi: <https://doi.org/10.1186/s40594-024-00510-6>
- [2] X. Chen and X. Wang, "Computational Thinking Training and Deep Learning Evaluation Model Construction Based on Scratch Modular Programming Course," *Comput Intell Neurosci*, vol. 2023, no. 1, p. 3760957, 2023, doi: <https://doi.org/10.1155/2023/3760957>.
- [3] S. Bhatt, K. Verbert, and W. Van Den Noortgate, "A Method for Developing Process-Based Assessments for Computational Thinking Tasks," *Journal of Learning Analytics*, vol. 11, no. 2, pp. 157–173, Aug. 2024, doi: 10.18608/jla.2024.8291.
- [4] B. Tan, H.-Y. Jin, and M. Cutumisu, "The applications of machine learning in computational thinking assessments: a scoping review," *Computer Science Education*, vol. 34, no. 2, pp. 193–221, 2024, doi: 10.1080/08993408.2023.2245687.
- [5] B. Azeddine, B. Khalid, "Exploring AI Tools and Large Language Models for Students' Performance Enhancement in Riddle Based Logical Reasoning", *International Journal of Modern Education and Computer Science (IJMECS)*, Vol.17, No.5, pp. 1-28, 2025. doi:10.5815/ijmees.2025.05.01
- [6] B. Palop, I. Díaz, L. J. Rodríguez-Muñiz, and J. J. Santaengracia, "Redefining computational thinking: A holistic framework and its implications for K-12 education," *Educ Inf Technol (Dordr)*, 2025, doi: 10.1007/s10639-024-13297-4.
- [7] R. Tariq, B. M. Aponte Babines, J. Ramirez, I. Alvarez-Icaza, and F. Naseer, "Computational thinking in STEM education: current state-of-the-art and future research directions," 2024, *Frontiers Media SA*. doi: 10.3389/fcomp.2024.1480404.
- [8] H. Zhang, J. Huang, Z. Li, M. Naik, and E. Xing, "Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming," May 2023, [Online]. Available: <http://arxiv.org/abs/2305.03742>
- [9] H. Hidayat, A. Komariah, B. B. Wiyono, and Y. Huda, "Impact of the Use of Fuzzy Comprehensive Evaluation Applications towards Computational Thinking Skill Students in Engineering Education," *International Journal of Information and Education Technology*, vol. 15, no. 1, pp. 90–100, 2025, doi: 10.18178/ijiet.2025.15.1.2221.
- [10] P.-H. Lin and S.-Y. Chen, "Design and Evaluation of a Deep Learning Recommendation Based Augmented Reality System for Teaching Programming and Computational Thinking," *IEEE Access*, vol. 8, pp. 45689–45699, 2020, doi: 10.1109/ACCESS.2020.2977679.
- [11] D. Li, K. D. Ortigas, and M. White, "Exploring the Computational Effects of Advanced Deep Neural Networks on Logical and Activity Learning for Enhanced Thinking Skills," *Systems 2023*, Vol. 11, Page 319, vol. 11, no. 7, p. 319, Jun. 2023, doi: 10.3390/SYSTEMS11070319.
- [12] S. Li, S. P. Lajoie, J. Zheng, H. Wu, and H. Cheng, "Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving," *Comput Educ*, vol. 163, p. 104114, 2021, doi: <https://doi.org/10.1016/j.compedu.2020.104114>.
- [13] A. Master, D. Tang, D. Forsythe, T. M. Alexander, S. Cheryan, and A. N. Meltzoff, "Gender equity and motivational readiness for computational thinking in early childhood," *Early Child Res Q*, vol. 64, pp. 242–254, Jul. 2023, doi: 10.1016/J.ECRESQ.2023.03.004.
- [14] L. Hu, "Exploring Gender Differences in Computational Thinking Among K-12 Students: A Meta-Analysis Investigating Influential Factors," *Journal of Educational Computing Research*, vol. 62, no. 5, pp. 1358–1384, Sep. 2024, doi: 10.1177/07356331241240670.
- [15] R. Richado, S. I. A. Dwiningrum, and A. Wijaya, "Computational Thinking Skill For Mathematics and Attitudes Based on Gender: Comparative and Relationship Analysis," *Pegem Journal of Education and Instruction*, vol. 13, no. 2, pp. 345–353, Feb. 2023, doi: 10.47750/PEGEGOG.13.02.38.
- [16] S. Jiang and G. K. W. Wong, "Exploring age and gender differences of computational thinkers in primary school: A developmental perspective," *J Comput Assist Learn*, vol. 38, no. 1, pp. 60–75, Feb. 2022, doi: 10.1111/JCAL.12591.
- [17] S. Lin and G. K. W. Wong, "Gender Differences in Computational Thinking Skills among Primary and Secondary School Students: A Systematic Review," *Education Sciences* 2024, Vol. 14, Page 790, vol. 14, no. 7, p. 790, Jul. 2024, doi: 10.3390/EDUCSCI14070790.
- [18] G. Ardito, B. Czerkawski, and L. Scollins, "Learning Computational Thinking Together: Effects of Gender Differences in Collaborative Middle School Robotics Program," *TechTrends*, vol. 64, no. 3, pp. 373–387, May 2020, doi: 10.1007/S11528-019-00461-8.
- [19] Q. Li, Q. Jiang, J. C. Liang, W. Xiong, and W. Zhao, "Engagement predicts computational thinking skills in unplugged activity: Analysis of gender differences," *Think Skills Creat*, vol. 52, p. 101537, Jun. 2024, doi: 10.1016/J.TSC.2024.101537.
- [20] S. Papavasiliou, K. Sharma, and M. N. Giannakos, "Coding activities for children: Coupling eye-tracking with qualitative data to investigate gender differences," *Comput Human Behav*, vol. 105, p. 105939, Apr. 2020, doi: 10.1016/J.CHB.2019.03.003.
- [21] L. García-Pérez, D. Roldán-Álvarez, and J. M. Cañas, "Does Gender Influence the Learning Process of Computational Thinking in Secondary Education?," *Computer Applications in Engineering Education*, vol. 33, no. 3, May 2025, doi: 10.1002/CAE.70050.

- [22] L. Sun and J. Liu, "Improving Computational Thinking for Middle School Students Through Python Programming: Interaction Effect Analysis of Grade Level and Gender," *Journal of Educational Computing Research*, vol. 63, no. 4, pp. 902–929, Jul. 2025, doi: 10.1177/07356331251332751.
- [23] A. M. Gasaymeh and R. AlMohtadi, "College of education students' perceptions of their computational thinking proficiency," *Front Educ (Lausanne)*, vol. 9, p. 1478666, Oct. 2024, doi: 10.3389/FEDUC.2024.1478666/BIBTEX.
- [24] J. Parraga-Alava and J. Rodas-Silva, "Computational Thinking and Academic Performance Across Different Instructional Modalities in Pre-University Courses: A Data-Driven Study," *Engineering Reports*, vol. 7, no. 6, Jun. 2025, doi: 10.1002/ENG2.70203.
- [25] A. De la Hoz Serrano, L. V. Melo Niño, A. Álvarez Murillo, M. Á. Martín Tardío, F. Cañada Cañada, and J. Cubero Juárez, "Analysis of Gender Issues in Computational Thinking Approach in Science and Mathematics Learning in Higher Education," *European Journal of Investigation in Health, Psychology and Education* 2024, Vol. 14, Pages 2865-2882, vol. 14, no. 11, pp. 2865–2882, Nov. 2024, doi: 10.3390/EJIHPE14110188.
- [26] Z. Demir-Kaymak, İ. Duman, C. Randler, and M. B. Horzum, "The Effect of Gender, Grade, Time and Chronotype on Computational Thinking: Longitudinal Study," *Informatics in Education*, vol. 21, no. 3, pp. 465–478, 2022, doi: 10.15388/INFEDU.2022.22.
- [27] Kwon, S. J. Lee, and K. Kim, "Gender disparities in K-12 computer science education: Status, contributing factors, and instructional approaches," *Computers and Education Open*, vol. 8, p. 100253, Jun. 2025, doi: 10.1016/J.CAEO.2025.100253.

Authors' Profiles



Prof. Azeddine Benelrhali is a doctoral researcher at Mohammed V University in Rabat, specializing in the intersection of Computational Thinking (CT), Artificial Intelligence in Education (AIED), and the United Nations Sustainable Development Goals (SDGs). He leads applied research on programming pedagogy, Open Educational Resources (OER), and AI-driven tools in educational practices across secondary and higher education, with a sustained focus on under-resourced and geographically marginalized learning contexts. His academic background combines an M.Sc. in High Energy Physics, Astronomy, and Computational Physics with advanced certifications in Intelligent Systems, Explainable AI (XAI), Innovative Education, and Data Analysis—a disciplinary breadth that bridges rigorous quantitative modeling and the applied learning sciences. His doctoral research deploys Large Language Models (LLMs) and Explainable AI (XAI) approaches to model and optimize student learning trajectories, with a specific focus on gender equity in STEM participation. Grounded in the realities of rural Moroccan communities, his work engages directly with SDG 4 (Quality Education), SDG 5 (Gender Equality), and SDG 10 (Reduced Inequalities), addressing the compounded barriers created by geographical marginalization, gender exclusion, and resource deficits to advance digital equity and spatial justice in education.



Dr. Khalid Berrada is a full professor of physics at Mohammed V University in Rabat, Morocco. Berrada was director of the Centre for Pedagogical Innovation at Cadi Ayyad University (UCA) and a UNESCO Chair in "Teaching Physics by Doing". He has participated in many national and international conferences and meeting committees. Berrada has published over 80 scientific papers, ten books and special issues in indexed journals. He is also one of the developers of the victorious French programme of UNESCO's Active Learning in Optics and Photonic and coordinated the UC@MOOC project in 2013 at UCA. Berrada has led a group of researchers on educational innovation at UCA (TransERIE) and the Morocco Declaration on Open Education since 2016. Berrada is currently the Director of the ICESCO Chair on Open Education (2023–2027). He is also the Director of Higher Education and Pedagogical Development at the Ministry of Higher Education, Scientific Research, and Innovation (2023-2025).

How to cite this paper:

Azeddine Benelrhali, Khalid Berrada, "Artificial Intelligence vs. Traditional Models in Computational Thinking Gap Analytics: A BERT-Driven Approach with XAI Diagnostics", *International Journal of Intelligent Systems and Applications (IJISA)*, Vol.18, No.5, pp.173-190, 2026. DOI:10.5815/ijisa.2026.05.09