machine learning & knowledge extraction

MDPI

Article

Augmenting Deep Neural Networks with Symbolic Educational Knowledge: Towards Trustworthy and Interpretable AI for Education

Danial Hooshyar [1,*], Roger Azevedo [2] and Yeongwook Yang [3]

[1] School of Digital Technologies, Tallinn University, 10120 Tallinn, Estonia
[2] School of Modeling Simulation and Training, University of Central Florida, Orlando, FL 32816, USA; roger.azevedo@ucf.edu
[3] Department of Computer Science and Engineering, Gangneung-Wonju National University, Wonju 26403, Republic of Korea; yeongwook.yang@gwnu.ac.kr
* Correspondence: danial.hooshyar@tlu.ee

Abstract: Artificial neural networks (ANNs) have proven to be among the most important artificial intelligence (AI) techniques in educational applications, providing adaptive educational services. However, their educational potential is limited in practice due to challenges such as the following: (i) the difficulties in incorporating symbolic educational knowledge (e.g., causal relationships and practitioners' knowledge) in their development, (ii) a propensity to learn and reflect biases, and (iii) a lack of interpretability. As education is classified as a 'high-risk' domain under recent regulatory frameworks like the EU AI Act—highlighting its influence on individual futures and discrimination risks—integrating educational insights into ANNs is essential. This ensures that AI applications adhere to essential educational restrictions and provide interpretable predictions. This research introduces NSAI, a neural-symbolic AI approach that integrates neural networks with knowledge representation and symbolic reasoning. It injects and extracts educational knowledge into and from deep neural networks to model learners' computational thinking, aiming to enhance personalized learning and develop computational thinking skills. Our findings revealed that the NSAI approach demonstrates better generalizability compared to deep neural networks trained on both original training data and data enriched by SMOTE and autoencoder methods. More importantly, we found that, unlike traditional deep neural networks, which mainly relied on spurious correlations in their predictions, the NSAI approach prioritizes the development of robust representations that accurately capture causal relationships between inputs and outputs. This focus significantly reduces the reinforcement of biases and prevents misleading correlations in the models. Furthermore, our research showed that the NSAI approach enables the extraction of rules from the trained network, facilitating interpretation and reasoning during the path to predictions, as well as refining the initial educational knowledge. These findings imply that neural-symbolic AI not only overcomes the limitations of ANNs in education but also holds broader potential for transforming educational practices and outcomes through trustworthy and interpretable applications.

Keywords: learner modeling; personalized learning; machine learning; knowledge extraction; neural-symbolic AI; explainable and trustworthy AI

Citation: Hooshyar, D.; Azevedo, R.; Yang, Y. Augmenting Deep Neural Networks with Symbolic Educational Knowledge: Towards Trustworthy and Interpretable AI for Education. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 593–618. https://doi.org/10.3390/make6010028

Academic Editor: Weiping Ding

Received: 26 January 2024
Revised: 29 February 2024
Accepted: 7 March 2024
Published: 10 March 2024

1. Introduction

Recent developments in artificial intelligence (AI) have paved the way for the widespread adoption of intelligent applications in diverse educational settings [1]. Intelligent educational systems incorporate a crucial element called a learner model, situated behind the user interface, which analyzes data on learner–system interactions [2]. This model, constructed through a computational process of learner modeling using AI techniques such as symbolic and sub-symbolic approaches, provides a structured representation of learners' knowledge

and learning states [3]. Consequently, it enables the delivery of personalized and optimal learning experiences, enhancing educational outcomes [4].

Symbolic AI methods rely on explicit representations and rules when processing information, offering the advantage of providing clear explanations and reasoning for AI decision making. However, these often have high costs in terms of human involvement and may struggle with real-world complexities and data quality issues [5,6]. In contrast, sub-symbolic AI operates with less reliance on explicit rules, demonstrating greater resilience to noisy and incomplete data. Sub-symbolic methods excel in tasks requiring pattern recognition and predictive performance, making them popular in various domains. The key distinction lies in the explicit representation of knowledge in symbolic AI versus the more implicit, data-driven approach of sub-symbolic methods. As a result, sub-symbolic methods like deep neural networks, which belong to the family of artificial neural networks (ANNs), have gained considerable popularity in various educational tasks, including learner modeling (e.g., [7–11]). Despite their success and popularity, they face three primary challenges that limit their educational value.

One of the main difficulties is incorporating educational constructs, restrictions, guidelines, causal relationships, and practitioners' knowledge (collectively referred to as *educational knowledge*) into their development. Deep neural networks primarily operate with numerical data, requiring the translation of all information into numerical values [12]. This becomes problematic when modeling learners' behavior and performance in digital learning due to the unavailability of representative training data or difficulty in collecting precise numerical data to capture the adaptations of complex and dynamic learning patterns. For example, translating complex educational constructs, such as student engagement [13], into numerical data, or integrating dynamic factors like student motivation into neural networks, is challenging due to their qualitative nature and the complexity of accurately translating these factors into numerical inputs. Although incorporating educational knowledge can enhance training data, its integration into deep neural network applications in education is still limited [14].

Secondly, like many other machine learning methods, deep neural networks are prone to learning misleading correlations during training, resulting in their dependence on irrelevant or unnatural features, which could limit their accuracy and generalizability [15]. This could lead to significant failures when deploying the model in real-world applications (e.g., [16–19]). In many cases, spurious correlations occur when a machine learning model relies on features that have a strong correlation with the target variable in the training data but are not causally related to it [20]. To ensure fairness in AI applications in education, it is crucial to avoid algorithmic bias and prevent algorithms from being tuned to favor a majority group solely to achieve high performance accuracy [1,21]. While recent research provides evidence that incorporating educational knowledge can enhance deep neural networks' understanding of training examples and reduce biases, such applications in education are extremely limited [22].

Finally, sub-symbolic methods, such as deep neural networks, struggle with explaining and providing reasoning for their decision-making processes. In this context, interpretability denotes the characteristics of a learning model that explains its decisions in a manner comprehensible to humans, offering insights into its underlying reasoning processes. In education, there is an urgent need to provide interpretability for several reasons [23,24]. Firstly, a lack of interpretability hampers trust in deep neural network applications for educators and students. Both teachers and learners require transparency to comprehend the rationale behind recommendations, assessments, or feedback. Secondly, interpretability is crucial for effective learning [25]. Students benefit from accurate and informative feedback. When AI systems cannot provide clear explanations for their recommendations or grading, students miss out on valuable learning opportunities. Thirdly, interpretability fosters the development of metacognitive and critical thinking skills. Education should nurture curiosity and encourage questioning and understanding of the information and decisions that are presented. Uninterpretable deep neural networks hinder this process,

limiting students' ability to question, analyze, or critique AI output. Lastly, the lack of interpretability can lead to bias and unfairness in educational AI systems. Without the ability to explain and understand the decision-making process, it becomes difficult to identify and rectify any biases that may exist in the data or the model itself. Recently, there have been several attempts to ensure the interpretability of deep neural network applications in education using common explainer algorithms like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) (e.g., [26–29]). While helpful in generating an approximation of the decision, such techniques face serious challenges (e.g., [30,31]). For instance, such explainer models can use a particular feature to make a prediction without that feature appearing in the explanation of the prediction, and they often produce unrealistic scenarios. Most importantly, such models are unable to provide the reasoning behind the path to their decisions.

One prospective means of addressing these challenges is a recently emerging paradigm in the AI research community that is called neural-symbolic AI or the third wave of AI [32,33]. By combining symbolic knowledge representation with deep neural networks, neural-symbolic AI offers a potential solution to enhance learner modeling in adaptive educational environments (e.g., [14,22,33,34]). This integration serves to leverage the interpretability and explanatory power of symbolic models while harnessing the scalability and predictive performance of deep neural networks. First, it augments the neural network with educational knowledge, compensating for the limitations of the training data by encoding guidelines and causal relationships as symbolic rules. This ensures that predictions are pedagogically sound, even when representative training data are lacking. Second, it prevents the AI from learning misleading correlations and biases from the data by grounding neural network predictions in validated educational principles rather than purely data-driven patterns. Third, the educational knowledge representation also acts as a foundation for the extraction of knowledge and rules from the trained network, enhancing its ability to explain its predictions. Consequently, neural-symbolic AI presents an opportunity to overcome the limitations of existing approaches and unlock new possibilities for more effective and efficient learner modeling [14]. Despite the evident potential, neural-symbolic AI lacks sufficient attention and research in the field of education. This study seeks to bridge this gap by introducing and developing a neural-symbolic AI approach, henceforth referred to as NSAI, to address the challenges facing the field of education. The NSAI approach models learners' computational thinking by incorporating both symbolic educational knowledge and training data in the construction of deep neural networks, and extracts hidden knowledge from the trained networks to explain the reasoning behind its predictions. We compare the performance of the NSAI with a multilayer perceptron (MLP) trained only on training data, as well as training data augmented by the Synthetic Minority Over-sampling Technique (SMOTE) [35] and an autoencoder [36]. To achieve this aim, we set the following research questions:

- How effectively can we ground educational knowledge in a deep neural-network-based learner modeling approach in order to provide control over the network's behavior?
- How is the performance of the NSAI approach, in terms of generalizability, in handling data biases, and how interpretable are the predictions compared to deep neural networks?
- What are the effects of the data augmentation methods of SMOTE and autoencoders on the prediction power of deep neural network models?

2. Related Works

2.1. Neural Networks in Education

Recently, the utilization of ANNs within various domains has experienced remarkable growth (e.g., [37–44]). In an educational context, this surge in applications includes diverse functionalities, such as predicting student performance, as demonstrated by the works of [9,10,45–47]. The predictive capabilities of ANNs extend to detecting undesirable student behavior, with contributions from Fei and Yeung [48], Teruel and Alemany [49], and Whitehill et al. [50]. Furthermore, ANNs have been instrumental in generating recommendations,

as evidenced by the studies of Abhinav et al. [51], Algarni and Sheldon [11], Bhanuse and Mal [7], and Wong [52]. For instance, Abhinav et al. [51] introduced a recommendation system that leverages ANNs for content-based filtering, alongside collaborative filtering techniques, to personalize learning opportunities. By analyzing learners' profiles and course histories and employing ANNs to extract features from textual data, the system predicts course ratings to offer personalized course recommendations. Wong [52] presents an automated course recommender system for higher education, utilizing Long Short-Term Memory (LSTM) neural networks to address the complexities of curriculum planning. It innovatively combines deep learning for sequential course recommendations with dynamic updates to accommodate sequence, constraints, and concept drift, offering personalized educational pathways tailored to individual student needs and evolving course content. In the realm of evaluations, ANNs have played a role in the research conducted by Hooshyar and Yang [53], Taghipour and Ng [54], and Zhao et al. [55]. For example, Hooshyar and Yang [53] proposed a novel image-based learner modeling approach using Convolutional Neural Networks (CNNs) and transfer learning to predict learners' performance by classifying their computational thinking solutions using image data. Zhao et al. [55] introduced an innovative automated grading model for essays and open-ended assignments, leveraging memory networks to handle the growing demand for Massive Online Courses (MOOCs). By collecting and utilizing graded samples for each score in a rubric as a benchmark, the model predicts scores for new submissions based on their similarity to stored responses. Beyond these specific examples, the comprehensive review by Hernández-Blanco et al. [8] provides a more detailed exploration of the myriad applications of ANNs in the educational context. Concerning student performance prediction, knowledge tracing that revolves around the prediction of students' future performance based on their past activities is a significant challenge in education. Initial attempts by Piech et al. [10] introduced deep neural network techniques that outperformed traditional machine learning, but their results faced scrutiny. Subsequent studies both supported (e.g., Wang et al. [46]) and challenged [56] Piech et al.'s work, comparing deep neural networks with traditional models and highlighting less significant differences. In a different context, deep neural network models were applied to analyze writing samples and clickstream data [57], improve knowledge retention [58], categorize learning capabilities (e.g., [59]), and develop a sequential event prediction algorithm [60]. These studies demonstrated the effectiveness of deep neural networks in handling large student datasets, outperforming traditional approaches.

Various works have addressed the detection of undesirable student behavior in education, focusing on three subtasks: dropout prediction in MOOC platforms, evaluating social functions, and student engagement in learning. Regarding dropout prediction, studies applied deep neural network techniques, achieving a superior performance compared to traditional machine learning methods. Dropout was defined and approached differently in the different works, with methods including classification, joint embedding, and personalized intervention models (e.g., [49,61]). To evaluate social function, Tato et al. [62] put forward a deep neural-network-based approach, evaluating essential social ability for adaptive social functioning. Concerning student engagement, deep neural network models that utilize various methods to assess student engagement have demonstrated high predictive accuracy (e.g., [63,64]). Furthermore, recommender systems utilizing deep neural networks were created to suggest learning opportunities and personalized paths according to students' preferences and individual requirements [51,52]. Despite their success, the majority of the mentioned works are incapable of ensuring that their developed deep neural network-based approach complies with educational knowledge and restrictions, do not properly take into account data biases (especially spurious correlations), underperform when it comes to datasets that are unrepresentative or small, and are unable to provide an explanation of their predictions to allow users to understand the reasoning behind their decisions.

2.2. Neural-Symbolic AI for Education

The paradigm of neural-symbolic AI has gained prominence in the AI/machine learning research community, as highlighted by Garcez et al. [5]. This approach actively combines experiential learning principles with reasoning based on acquired knowledge. Basically, neural-symbolic AI integrates symbolic and connectionist (sub-symbolic) paradigms by representing knowledge symbolically and employing neural networks for learning and reasoning processes [33]. Such integration allows for robust learning, logical reasoning, and interpretability.

Recent research has provided evidence of the effectiveness of this family of AI in addressing many existing challenges. For instance, Tran and Garcez [65] introduced and evaluated algorithms that could both insert and extract knowledge from deep networks. They inserted background knowledge into the weights of the neural networks through confidence rules to guide the learning process, resulting in improved generalizability and accuracy. In a similar vein, Hu et al. [66] combined deep neural networks with structured logic rules to enhance their flexibility and reduce the un-interpretability of neural models. Their proposed framework employs an iterative distillation method, transferring structured information from logic rules into the weights of neural networks. Their experimentation revealed the positive impact of such a knowledge infusion on network generalizability. Serafini and Garcez [67] introduced Logic Tensor Networks (LTNs), a theoretical framework integrating tensor neural networks with reasoning based on first-order many-valued/fuzzy logic. LTNs support various reasoning and learning tasks, combining a logical knowledge representation with data-driven machine learning. Their findings showed that LTNs could achieve improved performance through the utilization of background knowledge. Garcez et al. [5,33] and Yu et al. [68] provide surveys of neural-symbolic computing across various fields. Despite this extensive exploration in various domains, neural-symbolic AI lacks sufficient attention and studies in the field of education.

Recently, Hooshyar and Yang [14] introduced a framework merging neural-symbolic AI principles in educational contexts, emphasizing symbolic knowledge incorporation and providing interpretability. Moreover, Shakya et al. [69] proposed a neural-symbolic approach combining the semantics of symbolic models with recurrent neural networks like LSTM, demonstrating superior prediction accuracy for KDD EDM challenge datasets compared to HMMs and pure LSTM methods, with a focus on a smaller training data fraction. Hooshyar [22] introduced a novel neural-symbolic AI approach for learner modeling across time. This approach tracks learners' knowledge in a temporal manner by combining probabilistic graphical models (i.e., dynamic Bayesian networks) with unsupervised neural networks enhanced with educational knowledge. The results from cross-validation and practical applications show the effectiveness of the approach, achieving 85% accuracy and an 87% F1 score, surpassing dynamic Bayesian networks. The combination of augmented prediction and interpretable knowledge tracing promotes trust among educators and learners, highlighting the potential of neural-symbolic AI in creating trustworthy and interpretable applications in education. This research aims to extend the foundations laid by the previous research by proposing a neural-symbolic AI approach called NSAI, which not only injects educational knowledge but also transforms the model architecture. NSAI allows for the explicit integration of propositional educational knowledge during training, providing control over the model's behavior. Additionally, the NSAI approach addresses data bias issues, specifically spurious correlations, ensuring algorithmic fairness and compensating for the lack of training data and data inconsistencies. Finally, our approach enables the extraction of knowledge from trained deep neural networks, enhancing their applications with interpretability.

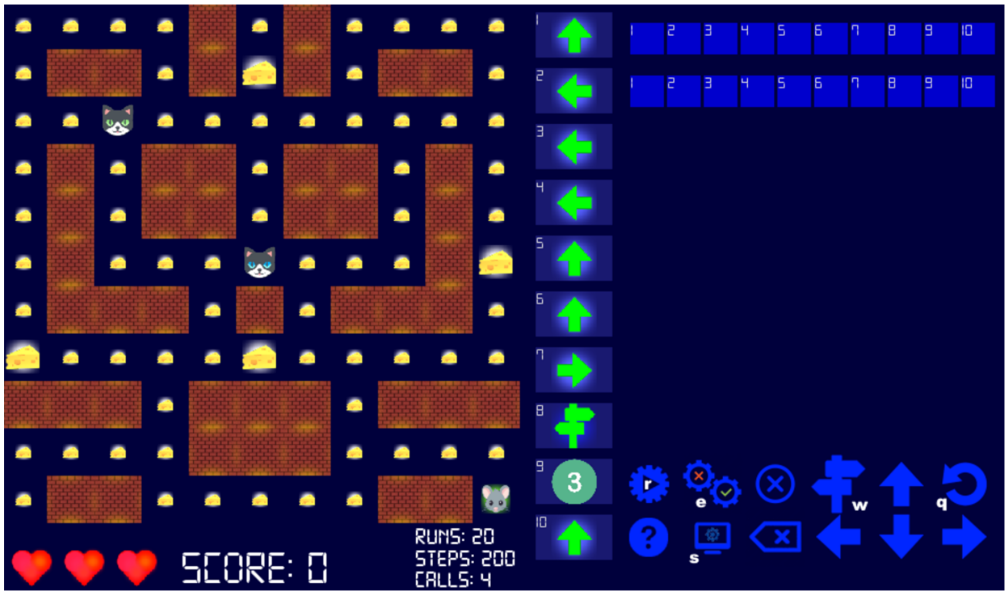
3. Neural-Symbolic AI for Modelling Learners' Computational Thinking

As we explore the application of neural-symbolic AI in modeling learners' computational thinking, it is crucial to ground our discussion to a practical context. To this end, we introduce the AutoThinking game, an adaptive educational game designed to

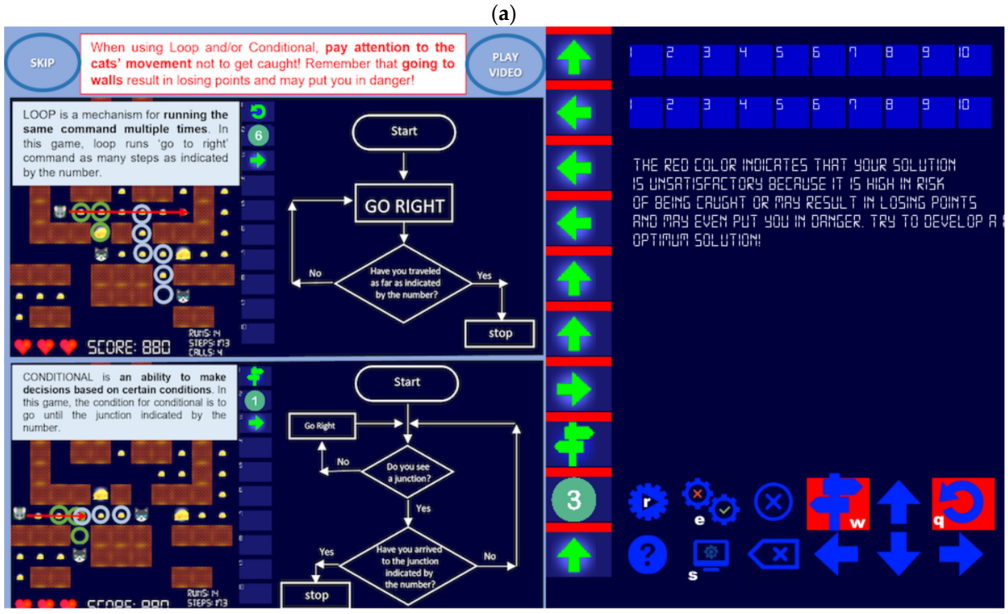
enhance computational thinking skills through interactive gameplay. This game serves as the primary data collection tool for our study, enabling us to gather rich insights into the decision-making processes and problem-solving strategies employed by learners.

3.1. The AutoThinking Game

AutoThinking is an adaptive educational computer game aimed at enhancing learners' computational thinking skills. It innovatively uses icons instead of traditional programming languages, minimizing syntax errors (for more details, see Hooshyar, Lim et al. [70]). In the game, players embody a mouse character as they collect cheese, earn points, and navigate a maze while avoiding cats. Players can devise up to 20 solutions to complete the game, with higher scores awarded for the effective application of computational thinking skills. The game fosters flexibility in developing solutions, incorporating functions to apply patterns in varied situations, and provides adaptive feedback and hints for guidance. Visual representations of a learner's solution, along with associated feedback and hints, are presented in Figure 1a,b.



(**a**)



(**b**)

Figure 1. (**a**) A learners' solution, and (**b**) generated feedback and a hint provided for the solution.

3.2. The Proposed NSAI Approach

To showcase the feasibility of applying NSAI frameworks to include educational knowledge, in addition to training data, in the development of (deep) neural networks, we adapted the KBANN framework developed by Towell and Shavlik [71]. In brief, this approach allows for the incorporation of symbolic knowledge into the architecture of neural networks to augment the training data, control the training flow of the network, and provide the context needed to map the learned representation into symbolic knowledge, which can interpret and provide reasoning for the decision making of the network. Table 1 presents the algorithm of the framework.

Table 1. Overall algorithm of the KBANN framework used in the proposed NSAI approach.

Step	Description
1	Rewrite rules (the symbolic knowledge in propositional logic form) to eliminate disjuncts.
2	Translate the rule structure into a neural network format.
3	Incorporate essential features that were not explicitly covered in the mapping.
4	Introduce hidden units into the architecture of the neural network.
5	Categorize units within the KBANN based on their respective levels.
6	Establish connections that are not explicitly defined by translating between all units in adjacent levels.
7	Introduce perturbations to the network by adding small random numbers to all link weights and biases.
8	Assign significant weight values to links derived from domain knowledge rules.
9	Apply backpropagation to refine the network to fit the training data.
10	Use weights and biases of the learned network to extract rules, explaining the predictions.

The NSAI approach begins by loading the training data and the educational knowledge in the form of rules (see datasets and the educational knowledge section). It then implements Towell's rewriting algorithm and, if there is more than one rule, this is rewritten as two rules [71]. It then establishes a mapping between the set of rules and the neural network to create layers, weights, and biases for the neural network. Finally, it applies backpropagation using training examples and uses the weights and biases of the learned network to extract rules, explaining the predictions.

Figure 2 illustrates the overall architecture of the NSAI approach. As shown in the figure, we first generated synthetical data to augment the training data using two methods of SMOTE upsampling and autoencoders (see the section focusing on setting up and evaluating the experiment). After that, we trained and evaluated the performance of the deep neural network using the three different sources: original data (called deep NN), original data augmented by SMOTE (called deep NN-SMOTE), and original data augmented by an autoencoder (called deep NN-Autoencoder). Additionally, we implemented the LIME method [72] to provide local and global explanations for the predictions. On the other hand, we developed the NSAI approach, which uses educational knowledge in addition to the original data. Upon training and evaluation, it also extracts rules from the networks to explain the predictions and provide the reasoning behind its decisions.

Hypothetical example: Let us assume that we have the domain knowledge and training examples presented in Table 2 related to self-regulated learning (SRL). For more information on the multilevel decomposition of the SRL strategies involved in multimedia learning, see Azevedo and Dever [73], Greene and Azevedo [74], Pekrun [75], and Winne and Azevedo [76].

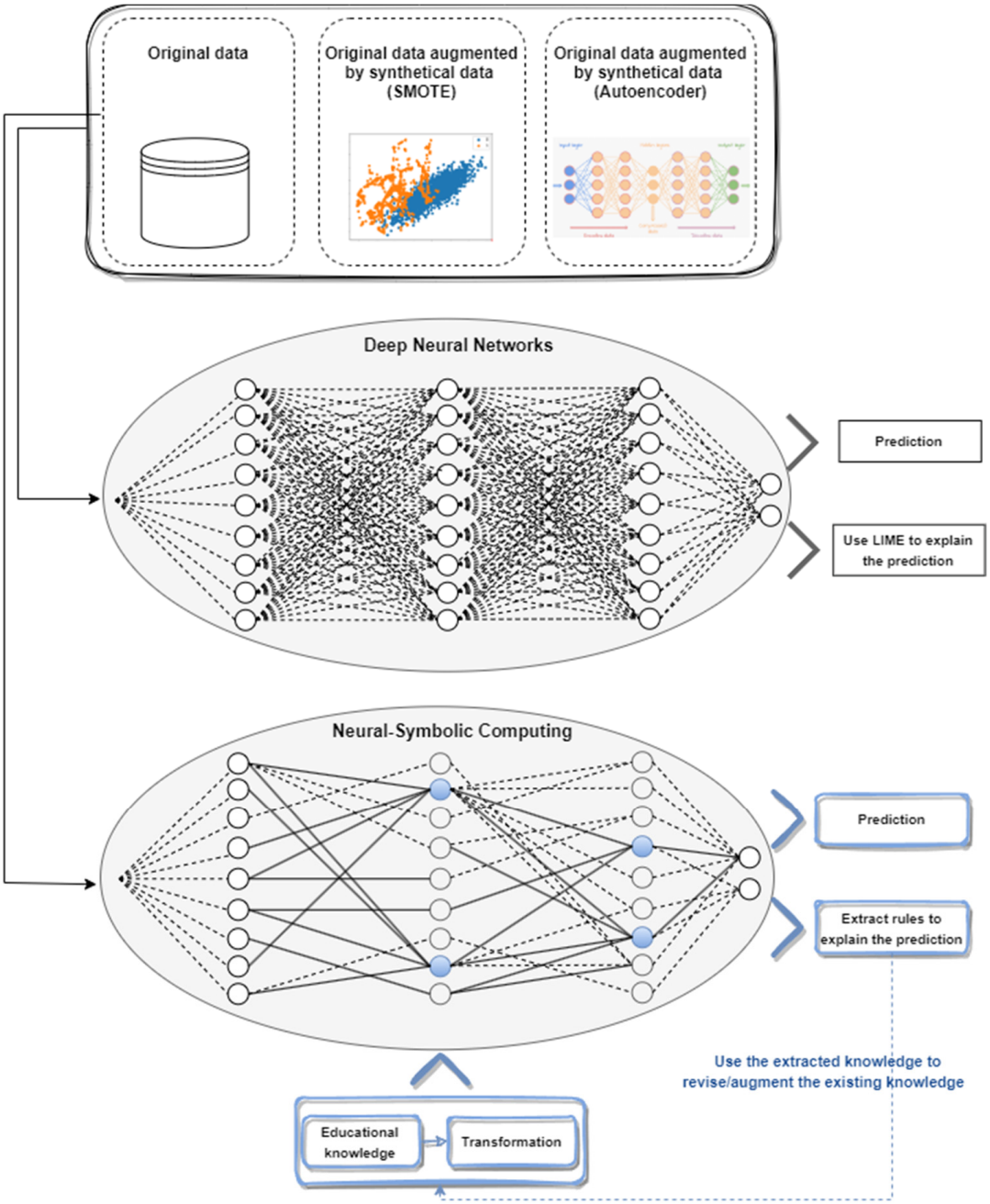


Figure 2. The overall architecture of the NSAI approach.

Symbolic domain knowledge:

Final performance: cognition, metacognition, emotion, motivation.
Cognition: planning, search for information, making inferences.
Metacognition: goal setting, information structuring, judgement of learning.
Information structuring: monitoring progress towards goals.
Emotion: seeking help, self-reported emotion rating.
Motivation: time spent watching learning materials, forum chat.

The first step is to employ the domain knowledge (which could be in the form of rules, logical relationships, etc.) to initialize the network architecture. This involves determining the number and types of layers, the number of neurons in each layer, and the connections between them. The architecture is constructed to reflect the problem-specific characteristics and constraints captured by the domain knowledge. Once the architecture is defined,

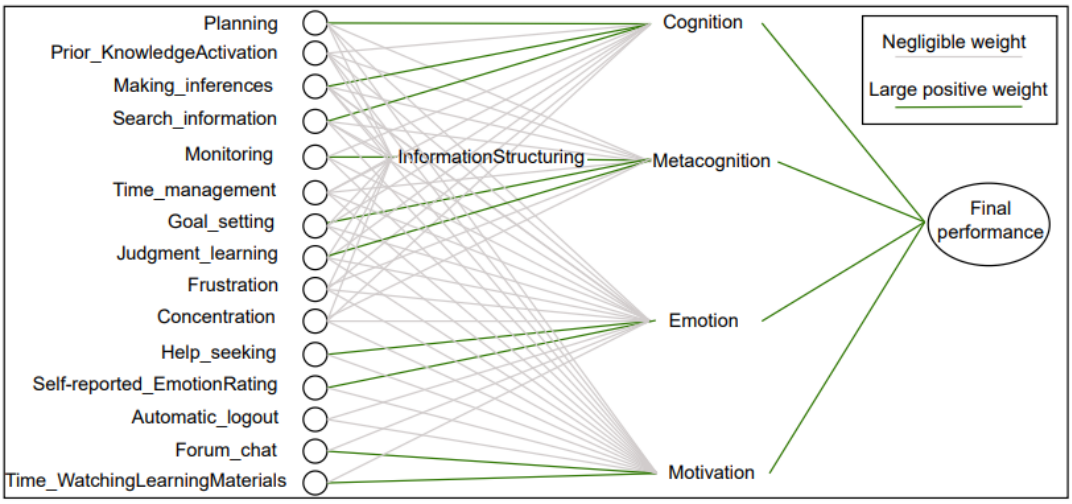
the domain knowledge is encoded into the network by incorporating the expert rules or constraints into the structure or parameters of the neural network. For example, certain connections may be fixed or constrained based on specific rules or relationships that were obtained from the domain knowledge. After initializing the network with the domain knowledge, the training process begins. Initially, the network's parameters (weights and biases) are typically randomly assigned or initialized. The training data, consisting of input–output pairs, are presented to the network. After this, forward propagation and error calculation take place, followed by backpropagation to adjust the network's parameters to reduce the overall error. This process repeats for multiple iterations or epochs, allowing the network to gradually improve its performance by learning from the training data and adjusting its parameters based on the error feedback. The training process continues until a termination criterion is met.

Table 2. Training examples.

Features	**Label (Final Performance)**									
	Low						**High**			
Goal Setting				✓			✓	✓	✓	
Prior Knowledge Activation			✓		✓					
Planning	✓			✓	✓	✓	✓	✓	✓	✓
Judgement of Learning	✓	✓			✓	✓	✓	✓	✓	✓
Time Management		✓					✓	✓	✓	
Monitoring (Progress Towards Goals)		✓	✓					✓		✓
Search for Information	✓					✓	✓			✓
Help Seeking		✓	✓		✓	✓				✓
Frustration	✓	✓				✓	✓		✓	
Time Spent Watching Learning Materials		✓	✓					✓	✓	
Self-reported Emotion Rating		✓					✓			
Concentration	✓		✓			✓	✓			✓
Forum Chat						✓	✓			
Automatic Logout		✓	✓	✓						
Making Inferences				✓			✓		✓	

In our example, as shown in Figure 3a, the network is initialized using the domain knowledge. This is shown using solid fixed connections between the SRL components and input features. For instance, the green line between *Cognition* and *Search for Information* indicates that learners with good information-searching skills during digital learning tend to have better cognitive skills. Similarly, for each rule in the domain knowledge, there is a fixed connection, with rather large weights between the features and the latent variable. Additionally, there are connections with negligible weights. These are connections that have very small weights, close to zero, after the initialization or during training. These connections may have a minimal impact on the network's output or learning process. After the initialization, the network is adjusted using the training examples. As Figure 3b shows, the initial network learned from the training examples and created a fixed connection with a high weight between *Metacognition* and *Time Management*. Moreover, it learned to create a fixed connection with large negative weights between *Motivation* and *Automatic Logout*, and *Emotion* and *Frustration*. This indicates that learners tend to have better affective states if they have a lower automatic logout and experience less frustration during digital learning (for more details, see [71]).

(**a**)



(**b**)

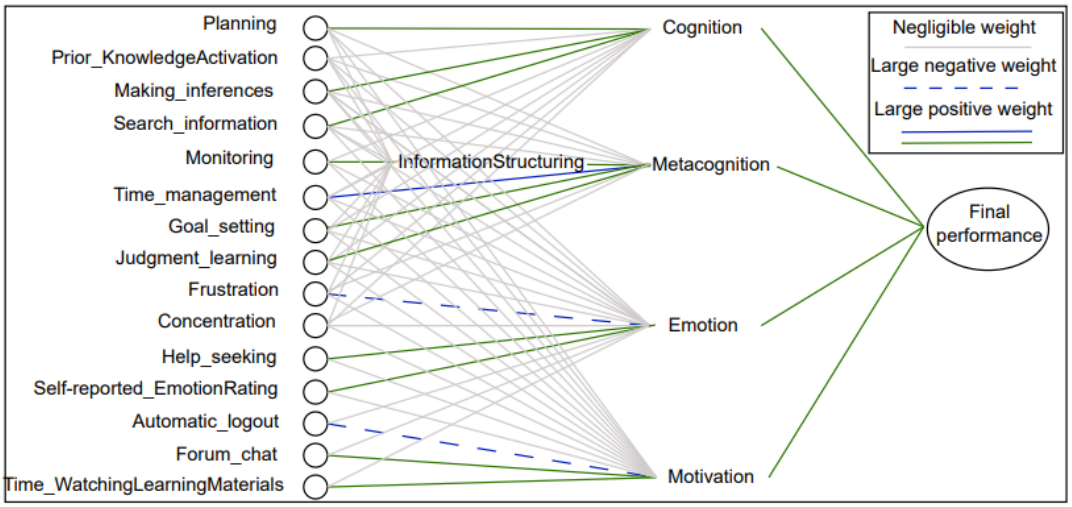


Figure 3. (**a**) Initializing the network using the domain knowledge, and (**b**) the adjusted network after training.

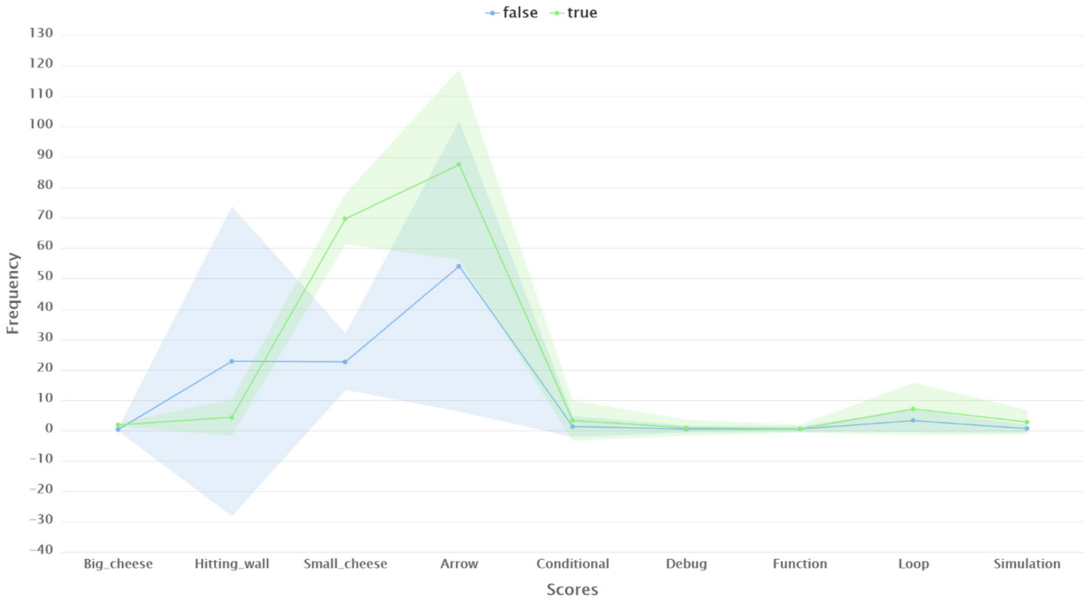
4. Results and Analysis

4.1. Datasets and Educational Knowledge

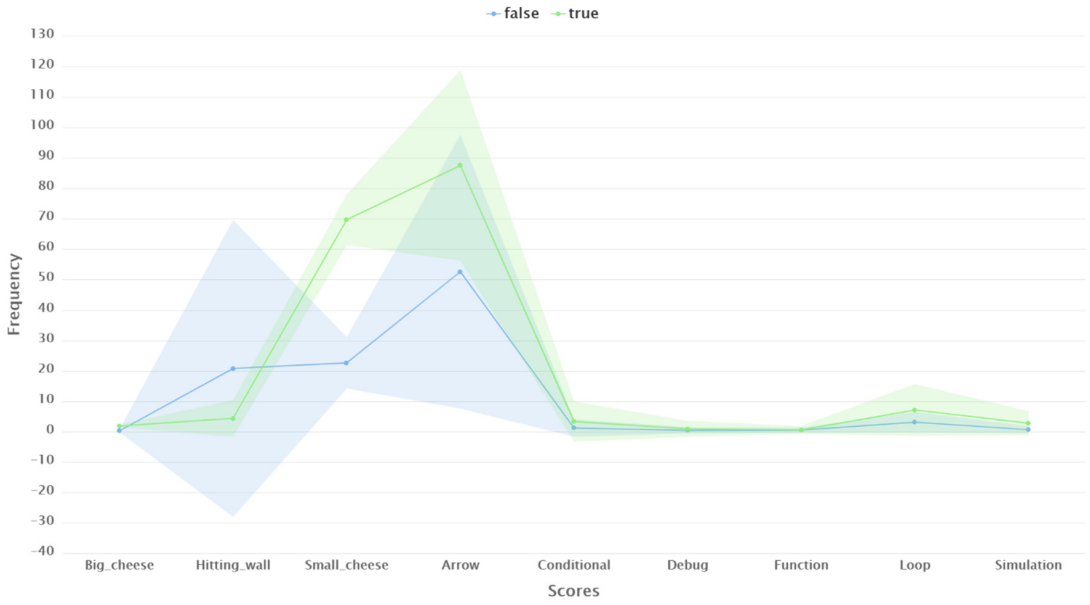
To model learners' computational thinking during gameplay, we utilized data from 427 players who participated in the third level of the AutoThinking game. During the gameplay, the system records different types of learner interactions. These include the following: (1) tracking the mouse and non-player characters' positions, (2) task identifiers, (3) the collection of small and large cheeses, (4) the usage of loops, conditionals, arrows, and functions, (5) debugging and simulation activity, (6) the frequency of seeking help, (7) the amount of feedback and hints that were received, (8) the frequency of colliding with walls, (9) the estimation of the learner's CT knowledge, and (10) the evaluation of solution quality inferred from the Bayesian network decision-making algorithm employed in the game. More information regarding the decision-making process is given in the work by Hooshyar, Lim et al. [70]. For this research, we selected the features that are directly/indirectly related to players' performance. Table 3 and Figure 4 provide a summary and distribution of the datasets, respectively.

Table 3. Distribution of the dataset before splitting.

Features	Min	Max	Average	Deviation
Arrow	15	180	82.05	34.65
Big cheese	0	4	1.6	0.7
Small cheese	0	74	63.38	17.72
Function	0	4	0.6	1.2
Debug	0	17	0.8	2.3
Simulation	0	19	2.92	4.24
Loop	0	50	6.66	8.12
Conditional	0	46	3	6.4
Hitting wall	0	180	6.19	18.57
Final score	True = 364		False = 63	

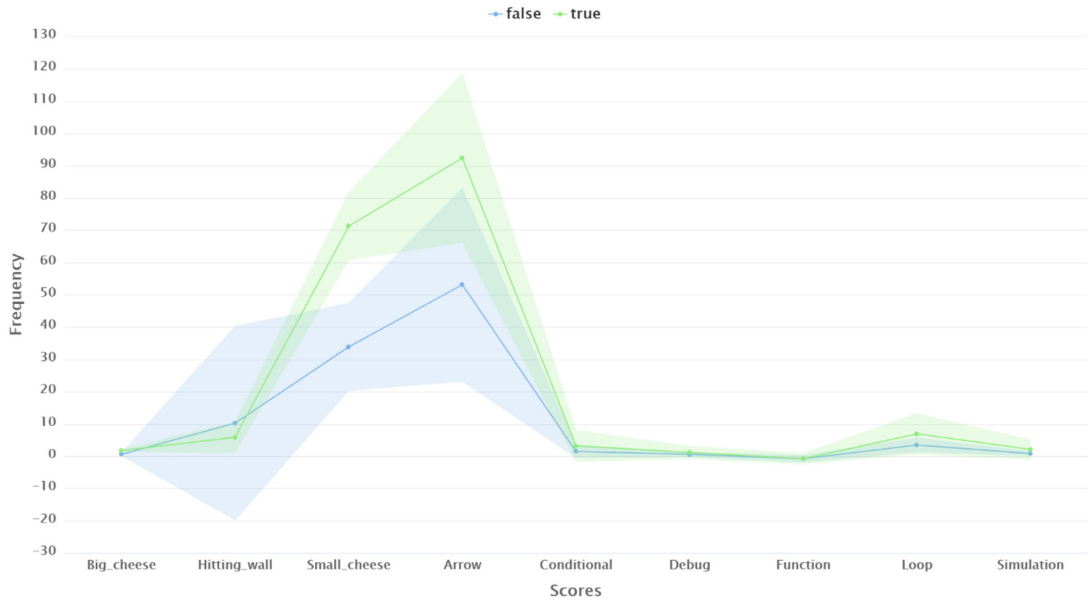


(**a**)

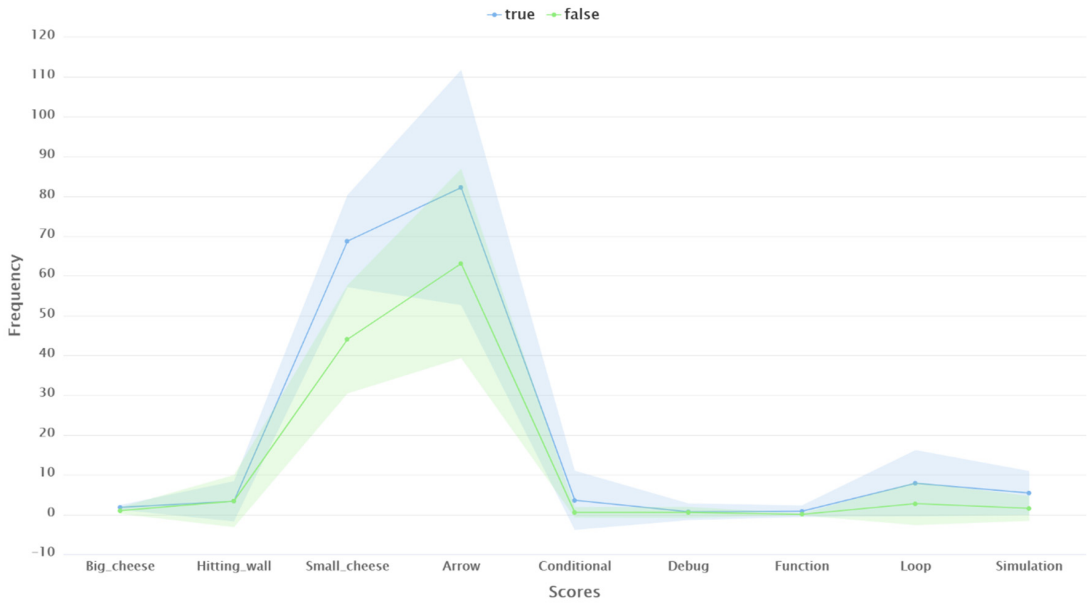


(**b**)

Figure 4. *Cont.*



(**c**)



(**d**)

Figure 4. Deviation chart showing the distribution of (**a**) the original training, (**b**) training augmented with SMOTE, (**c**) training augmented with autoencoder, and (**d**) test data (split using stratified sampling).

To introduce the domain knowledge into the NSAI approach, we adapted the CT skills and concepts used by Hooshyar [77]. This includes the CT skills of problem identification, building algorithms, debugging, and simulation, as well as CT concepts of sequences, loops, and conditional. To present the knowledge in the neural networks, we used propositional logic (i.e., a set of propositional, non-recursive Horn clauses), as shown in Table 4. In brief, the first propositional logic indicates that the final score or performance in the game depends on the players' mastery of CT skills and concepts (for more details, see [77]). Additionally, it states that the CT concepts are associated with conditional and loop features in the training dataset, whereas the CT skills are dependent on the debug, simulation, and function features. In other words, learners' overall performance in CT is good if they are good at CT concepts and skills.

Table 4. Educational knowledge represented in the symbolic form of rules.

Rule No.	Propositional Logic Representation	Explanation
1	Final_score: CT_concepts, CT_skills.	Final score or performance in a given task is contingent on achieving mastery in both CT skills and concepts.
2	CT_concepts: Conditional, Loop.	CT concepts are linked to the features of conditional and loop structures within the training dataset.
3	CT_skills: Debug, Simulation, Function.	CT skills rely on proficiency in debugging, simulation, and function features.

4.2. Data Biases

In the context of a training dataset, data bias refers to the presence of systematic and non-random errors or distortions in the data that can impact the performance and fairness of machine learning models. This occurs when certain subsets or categories within the dataset are overrepresented or underrepresented, leading to a skewed understanding of the underlying patterns and relationships. When a dataset is biased, it fails to accurately reflect the intended application of a model, leading to distorted results, reduced accuracy, and analytical mistakes [20].

Data bias in a training dataset can arise due to various factors, such as the sampling process, data collection methods, and presence of confounding variables (European Union Agency for Fundamental Rights [78]). These issues can impact the performance and fairness of machine learning models. Recently, there has been some research studying data biases and their effect on the performance of machine learning models. For instance, Blodgett et al.'s [79] work mainly revolves around ensuring that models do not rely on sensitive features such as race and gender, and Johnson and Khoshgoftaar [80] focus on addressing class imbalances and skewed distributions. However, when it comes to education, despite its classification as a high-risk domain by regulatory frameworks like the EU AI Act (https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf, accessed on 15 January 2024) due to its significant impact on individuals' futures and the inherent potential for perpetuating biases, as well as the widespread use of ANNs, there is not much research taking into account spurious correlations biases and their effect on the performance of machine learning (especially ANNs). Spurious correlations pose a major challenge when deploying machine learning models because they can cause the models to depend on irrelevant or unnatural features, leading to significant failures when deploying the model in real-world applications (e.g., [16,17,19]).

In many cases, spurious correlations can occur when a machine learning model relies on features that have a strong correlation with the target variable in the training data but are not causally related to it. For instance, in sentiment classification, bias can arise in the training data where positive examples tend to be longer than negative examples. In such cases, the model may erroneously consider length as a crucial feature for classification, even though it is a spurious feature that does not provide sentiment information [81]. In the case of the AutoThinking game, an example could be the existence of a strong correlation between the input variable of Small_cheese and the label (Final_score). Although the inclusion of the Small_cheese feature in the model could offer some useful information, it is important to note that there is no causal relationship between this feature and the class label. This is because many players employ random strategies in the game, where their solution can be successful in collecting both small and big cheeses without necessarily utilizing the main CT skills and concepts. For instance, a solution that solely uses arrows to navigate the game, without the NPC catching the mouse, can still collect many Small_cheese features and accordingly achieve high scores. In educational games, a high score can be obtained through either a random strategy or an appropriate strategy, such as parallel thinking [82].

Consequently, Tables 3 and 5 indicate the presence of class imbalance and potential data biases, respectively, caused by spurious correlations in the training dataset.

Table 5. Correlation between features and class labels in training and test data.

Features	Final Score			
	Training Data	SMOTE-Augmented Training Data	Autoencoder-Augmented Training Data	Test Data
Arrow	0.322 [1]	0.412	0.51	0.255
Big_cheese	0.728	0.858	0.68	0.445
Conditional	0.107	0.203	0.154	0.175
Debug	0.066	0.124	0.135	0.027
Function	0.011	0.027	0.005	0.233
Hitting_wall	−0.31	−0.23	−0.122	−0.003
Loop	0.164	0.295	0.239	0.25
Simulation	0.194	0.343	0.189	0.284
Small_cheese	0.887	0.942	0.807	0.632

[1] Red and blue represent the lowest and highest correlation values, respectively.

4.3. Experiment Setting and Evaluation

The computer utilized for our implementation featured a single AMD Ryzen 5 PRO 4650U CPU with 16.0 GB of memory. The deep learning model was structured as a multi-layer, feed-forward artificial neural network, which underwent training via stochastic gradient descent with back-propagation. A learning rate of 0.03 was chosen, and the Adam optimizer was employed. To prevent overfitting, we employed early stopping with a strategy based on score improvement, and a patience of three. Additionally, we applied regularization with a value of 1 for both L1 and L2. The model consisted of two fully connected layers with ReLU activation functions and 50 neurons, followed by an output layer with Softmax activation and two neurons. To evaluate our approach's performance, we utilized metrics such as accuracy, recall, and precision. Aside from the model trained only on the training dataset, we also augmented the training dataset using SMOTE upsampling and autoencoder methods, equalizing the classes. A similar feed-forward architecture to the autoencoder used to generate synthetical data was employed to augment the training dataset. The learning rate, optimizer, and regularization values were set to 0.03, Adam, and a value of one for both L1 and L2, respectively. Similar to the trained deep neural network, we employed early stopping with a strategy based on improving the score. The encoder and decoder employed ReLU activation functions. The encoder consists of three fully connected layers with eight, four, and two neurons, while the decoder comprises three layers with four, eight, and ten neurons in the last layer, utilizing the mean square error loss function. To evaluate the deep neural network models, we employed 10-fold cross-validation and tested the model on the test dataset.

4.4. Performance of Models in Terms of Generalizability

Table 6 provides a summary of the performances of the models using different metrics. While the differences in the overall performance metrics are modest, the NSAI model demonstrates a slight edge in terms of its generalizability, which is particularly notable in low-performance scenarios, where it shows a more substantial improvement compared to other models. In contrast, the deep NN-SMOTE model exhibits the least effectiveness across the evaluated metrics.

Table 6. Performance of models using test data from various learning sources.

Models	Source of Learning	Accuracy (%)	Recall (%)		Precision (%)	
			High	Low	High	Low
Deep NN	Training data	83.53	85.81	75.00	93.65	54.55
Deep NN-SMOTE	Training data + synthetical data	82.35	85.51	68.75	92.19	52.38
Deep NN-Autoencoder	Training data + synthetical data	83.53	86.00	68.75	92.31	55.00
NSAI	Training data + educational knowledge	84.71	86.00	81.00	95.00	57.00

Specifically, the NSAI model could achieve an accuracy of almost 85% and a recall of 86% and 81% for the *high* and *low* performers on the unseen data. This emphasizes that the NSAI model not only exhibits the highest likelihood of accurately recognizing a substantial portion of *high*-performing individuals, but also outperforms other models in effectively identifying learners with *low* performance. Similarly, the NSAI model has also been shown to have the best performance regarding the precision of *high-performer* learners (of all learners predicted to be *high* performers, 95% of the predictions were correct), and it is ranked the best concerning the precision of *low* performers. Given the critical importance of correctly identifying *low*-performing students in the education domain, the NSAI model's high recall rates make it a valuable asset, minimizing the risk of negative consequences for students' future educational outcomes and overall development.

The second-best model in terms of generalizability is the deep NN model, which is trained on original training data. This model exhibits a high ability to correctly predict *high* performers (recall of nearly 86%) out of all actual samples. However, when it comes to identifying *low* performers, its performance is relatively weaker, classifying only 75% of them out of all actual samples. The deep NN model trained on original training data augmented by SMOTE also exhibits the poorest performance in terms of generalizability. In practical scenarios, this model is unable to accurately identify *low*-performing learners, achieving a classification rate of less than 70%. Interestingly, all the deep NN models that were trained on training data (original or both original and synthetical) appeared to face a drop in performance in terms of their generalizability. However, when evaluating the NSAI model's performance on unseen data, it becomes evident that it has better generalizability due to its learning from both training data and explicit knowledge. The incorporation of explicit knowledge in the NSAI model provides it with a deeper understanding of the underlying relationships among training examples. This enables the model to have more effective learning even in situations where the training dataset is unrepresentative or when the test set exhibits a different distribution compared to the training set. By leveraging explicit knowledge, the NSAI model gains an advantage in its ability to generalize beyond the specific characteristics of the training data. This is particularly beneficial when faced with new or unseen data, as it can draw upon its broader understanding of the domain and the educational context.

Finally, regarding the effectiveness of augmenting the training data, while autoencoder augmented data appear to slightly improve some aspects of the model performance in terms of generalizability (e.g., recall of *high* performers), it causes a decrement in other aspects, like the recall of *low* performers. Interestingly, the SMOTE method even resulted in a slight drop in performance compared to the original training.

4.5. Performance of Models in Terms of Handling Data Biases and the Interpretability of Predictions

To determine if trained models learned biases from the training data and incorporated them into their predictions, we employed a two-step approach. Firstly, we conducted

a correlation analysis to assess the presence of strong positive or negative correlations between the features and the class label (investigating potential spurious correlations). While this analysis alone cannot definitively establish whether the models genuinely learned to rely on features that are not causally related to the target variable but are strongly correlated with it in the training data or disregarded the existing relationships and patterns among other features during its decision making, examining the internal workings of the models can complement the analysis and provide insights into whether biases were learned and reflected. If the prediction explanation highlights the spurious correlations as the most influential reasons for the models' decision-making process, and we discover that this is the reason for the models' misprediction of the test data, we can conclude that the models failed to grasp certain underlying patterns and relationships due to the presence of biases (represented by the spurious correlations).

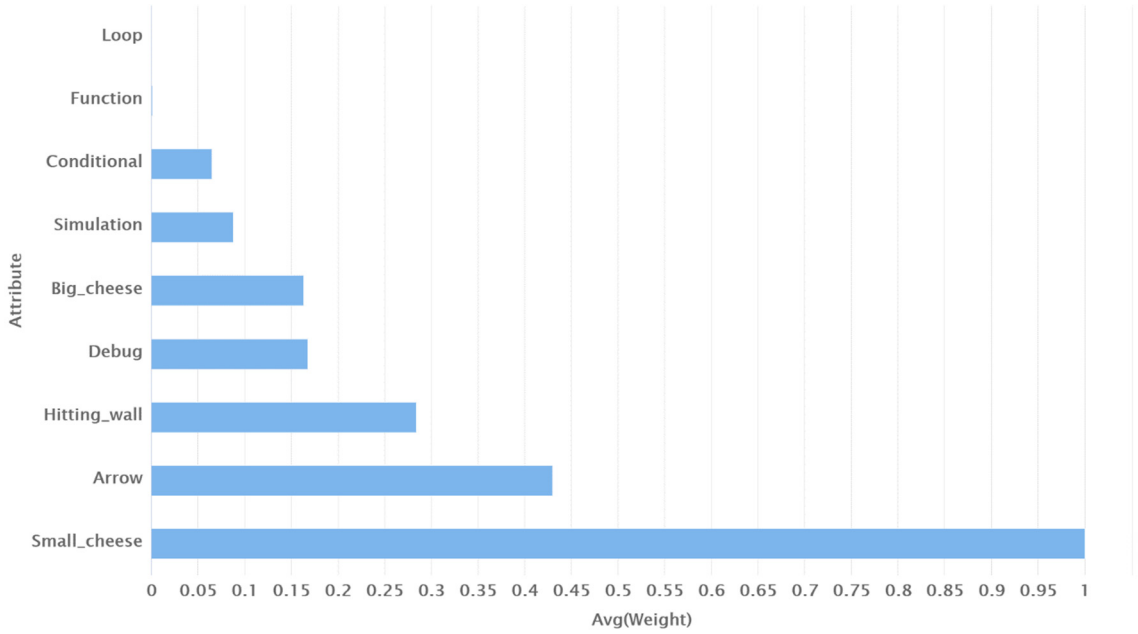
The results of the correlation analysis of both training and test data are shown in Table 4. While most features are positively correlated to the label, Hitting_wall has a negative correlation with the label, meaning that the less often the players bump into the walls, the higher their final performance. Amongst the positive correlations, in the training data, there is a strong positive correlation between consuming Small_cheese with the Final_score. Specifically, the Small_cheese feature has a positive correlation of 0.887 with the class label. This can cause model learning biases based on this spurious correlation, where models disregard the importance of other features and their relationships to the label. As mentioned previously, while the inclusion of the Small_cheese feature could offer some useful information, there is no causal relationship between this feature and the class label, and higher small cheese consumption does not necessarily indicate better CT concepts and skills. When it comes to the test data, while the direction of the correlation is similar, the strengths are weaker compared to the training data. Consequently, the models may heavily rely on features with strong spurious correlations during training, limiting their generalizability to test data that lack such strong correlations in their distribution.

To further investigate this matter and provide interpretability to the predictions, we implemented the LIME method in the first three models and extracted rules from the NSAI model.

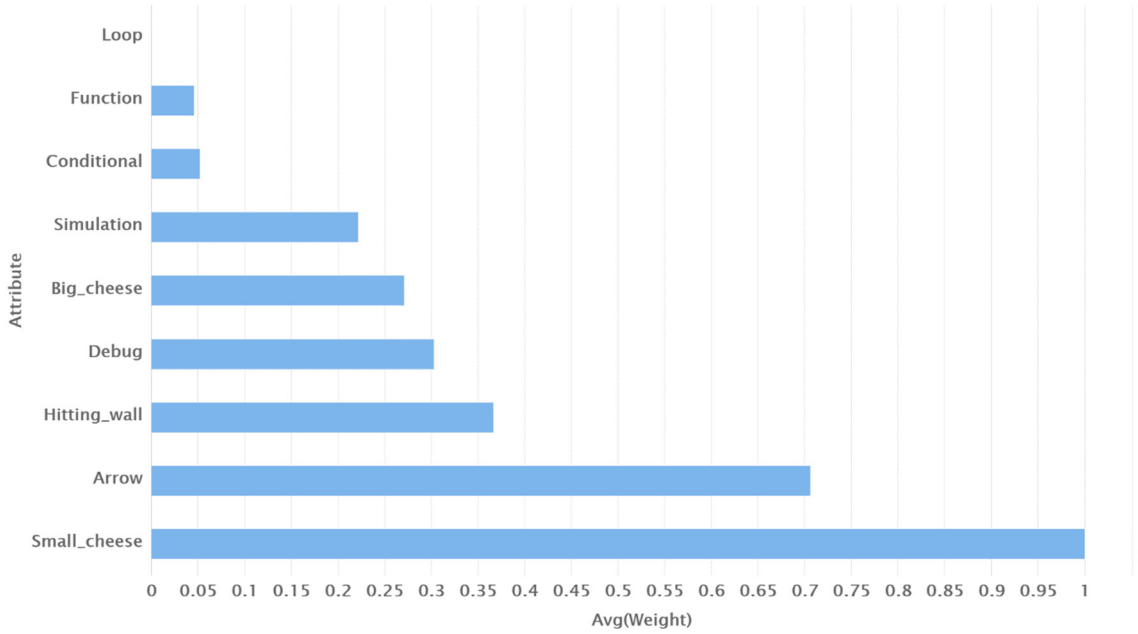
Figure 5 presents LIME explanations for the model predictions. As can be seen from Figure 5a, the deep NN model learned to heavily rely on features that are not causally related to the target variable but are strongly correlated with it (i.e., the small cheese). Moreover, it gives a higher weight to features like Arrow and Hitting Wall, along with Big Cheese consumption, which are not explicit predictors/indicators of CT knowledge. More importantly, the model fully ignored the Loop and Function features and paid little attention to Conditional and Simulation features, which are all causally related to CT skills and concepts. Similarly, both the deep NN-SMOTE and deep NN-Autoencoder used Small_cheese, followed by Arrow, as the primary features in their decision-making process, fully ruling out the important feature of Loop, and paid little attention to the crucial features of Function and Conditional in their decision making. Consequently, while the models employed a combination of features in their predictions, the feature with the highest correlations appears to be the primary feature in the models' decision making, and the models fail to consider some essential features, like Loop, Function, and Conditional, during the testing stage.

Table 7 lists three examples in which the models mispredicted the examples during the testing. The models incorrectly predicted a final score of *low* as *high* with full confidence due to the high levels of cheese consumption (after learning the spurious correlations). For instance, the first row of the table shows that the deep NN model incorrectly predicted a *low-performer* learner as *high*, mainly because of their high usage of Small Cheese and Arrow, which are not causal predictors of CT knowledge (allocating almost two-thirds of the entire feature importance to these two features). Considering the results of the correlation analysis and the global and local LIME explanations, it could be concluded that the models heavily relied on the spurious correlations learned during training and ignored

important factors, limiting their generalizability to test data. In other words, the models learned and reflected biases in their decision making by learning spurious correlations that caused them to overlook essential/underlying information.

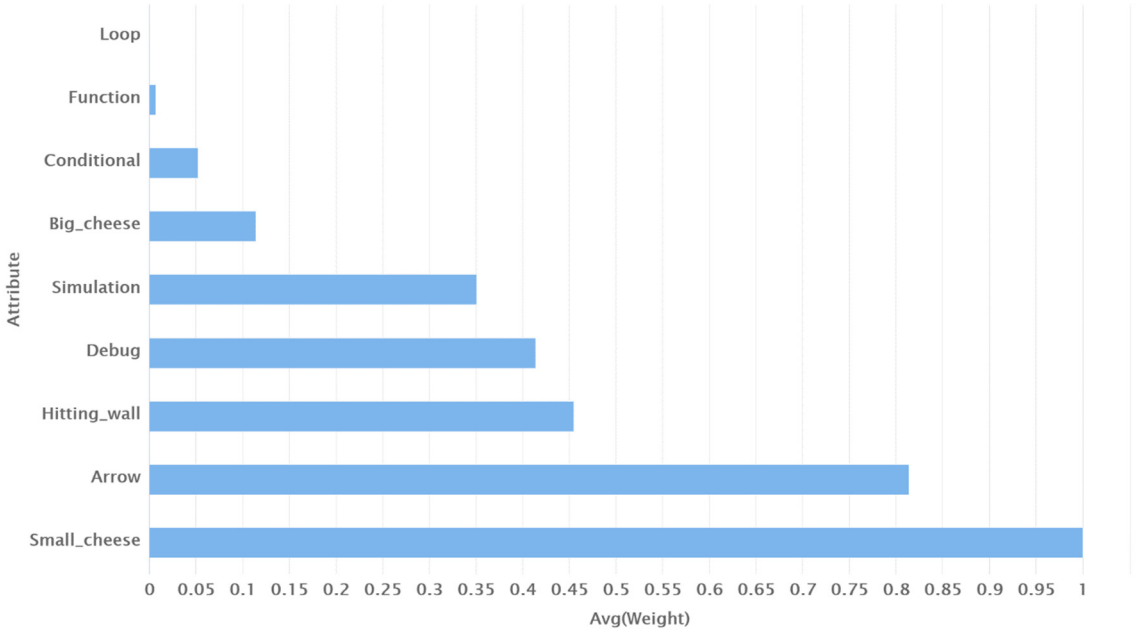


(**a**)



(**b**)

Figure 5. *Cont.*



(**c**)

Figure 5. Global LIME explanations: (**a**) deep NN, (**b**) deep NN-SMOTE, and (**c**) deep NN-Autoencoder.

Table 7. Local LIME explanations for some examples of mispredicted cases in the test dataset.

Model	Ground Truth	Prediction	Confidence (Low, High)	Supporting	Contradicting
Deep NN	Low	High	0.000, 1.000	Small_cheese = (value = 70, importance = 0.465) Arrow = (value = 76, importance = 0.204)	Hitting_wall = (value = 9, importance = −0.581) Conditional = (value = 0, importance = −0.162)
	Low	High	0.000, 1.000	Small_cheese = (value = 61, importance = 0.463) Arrow = (value = 91, importance = 0.204)	Hitting_wall = (value = 2, importance = −0.582) Conditional = (value = 2, importance = −0.162)
	Low	High	0.000, 1.000	Small_cheese = (value = 67, importance = 0.463) Arrow = (value = 100, importance = 0.203)	Hitting_wall = (value = 4, importance = −0.582) Conditional = (value = 0, importance = −0.163)
Deep NN-SMOTE	Low	High	0.000, 1.000	Small_cheese = (value = 70, importance = 0.397) Arrow = (value = 76, importance = 0.282)	Hitting_wall = (value = 9, importance = −0.643) Conditional = (value = 0, importance = −0.109)
	Low	High	0.000, 1.000	Small_cheese = (value = 61, importance = 0. 397) Arrow = (value = 91, importance = 0.282)	Hitting_wall = (value = 2, importance = −0. 643) Conditional = (value = 2, importance = −0.109)
	Low	High	0.000, 1.000	Small_cheese = (value = 67, importance = 0. 397) Arrow = (value = 100, importance = 0.282)	Hitting_wall = (value = 4, importance = −0. 643) Conditional = (value = 0, importance = −0.109)

Table 7. *Cont.*

Model	Ground Truth	Prediction	Confidence (Low, High)	Supporting	Contradicting
Deep NN-Autoencoder	Low	High	0.000, 1.000	Small_cheese = (value = 61, importance = 0.339) Arrow = (value = 91, importance = 0.290)	Hitting_wall = (value = 2, importance = −0.683) Conditional = (value = 2, importance = −0.094)
	Low	High	0.000, 1.000	Small_cheese = (value = 67, importance = 0.339) Arrow = (value = 100, importance = 0.290)	Hitting_wall = (value = 4, importance = −0.683) Conditional = (value = 0, importance = −0.094)
	Low	High	0.000, 1.000	Small_cheese = (value = 59, importance = 0.339) Arrow = (value = 86, importance = 0.290)	Hitting_wall = (value = 0, importance = −0.683) Conditional = (value = 0, importance = −0.094)

The rules extracted from the NSAI approach are shown in Table 8. As can be seen, unlike the model explanation provided by the LIME model, not only does the NSAI approach provide a combination of (observable and unobservable) features that contributed to the model's final decision-making process, but it also highlights the learned representations (i.e., latent features) and features contributing to the model's estimations/construction. More explicitly, as the first row of the table shows, learners' final performance/score in CT is predicted to be *high* provided that the value of the combination of learned representations of CT concepts and skills, as well as heads (in the NSAI, heads refer to the individual neural network nodes or units within a layer; specifically, each head represents a single neuron in a network layer), is larger than the threshold of 4.64. In other words, learners' final CT performance is predicted to be *high* if they are good at CT skills and concepts, as well as other heads that are related to learners' performance concerning other skills. Given that CT skills (with a weight of 2.00) and concepts (with a weight of 0.83) are both taken into account in the decision making, along with a combination of supporting (e.g., Small and Big Cheeses, Hitting the Wall) and contradictory features (e.g., Conditional, Loop, Arrow, Debug, Simulation) indicates that the model properly considers the skills and concept of CT in its decision making and complies with the introduced causal relationships. Furthermore, in line with the knowledge added to the model structure during the training, the learned representations of CT concepts mainly consider the Conditional and Loop features (with a weight of 3.00) in their estimations, whereas CT skills mainly consider the Debug, Function, and Simulation features (see the second and third rows of Table 7). Finally, heads 1, 2, and 3 are constructed according to a combination of supporting (e.g., Small and Big Cheeses, Hitting Wall, Simulation) and contradictory features (e.g., Conditional, Loop, Arrow, Debug). Consequently, the NSAI model clearly takes into account the causal relationships related to CT skills and concepts (the educational knowledge presented in the form of rules), and accordingly earns the underlying patterns from the training data. This implies that, unlike the other three models, which mostly focus on small cheese consumption and learn spurious correlations to improve the accuracy of the model when using test data, the NSAI model took into account educational knowledge and adhered to educational restrictions while learning from the data to improve accuracy regarding the test data. It is worth noting that the extracted rules also allow for the initial educational knowledge to be revisited using the training data. In this way, CT skills play a more important role in predicting the final performance of learners compared to CT concepts (see the weights associated with CT skills and concepts in the first row of the table).

Table 8. Model explanation for the NSAI using rule extraction from neural networks.

Variables	Rules
Final_score:	4.6377187 < 2.4166102 × (head 2, head 3) + 0.8252018 × (CT_concepts) + 2.0046637 × (CT_skills) + 1.7674259 × (head 1)
CT_concepts:	4.6082096 < 0.2453651 × (Small_cheese) + 3.002934 × (Conditional, Loop) + 0.0472862 × (Debug, Simulation, Function, Big_cheese, Hitting_wall) + −0.07132121 × (Arrow)
CT_skills:	8.519821 < 0.20699154 × (Small_cheese) + 2.3430111 × (Simulation) + 1.0791004 × (Function) + −0.18917799 × (Conditional, Loop) + 2.6324146 × (Debug) + 0.45198494 × (Big_cheese) + −0.0066499244 × (Arrow) + −0.11537525 × (Hitting_wall)
head 1:	2.2751489 < −0.070589505 × (Conditional, Loop, Debug, Arrow) + 0.80795884 × nt(Big_cheese) + 0.2296475 × (Hitting_wall) + −0.43813425 × (Function) + 0.09194418 × (Small_cheese) + 0.0072831404 × (Simulation)
head 2:	2.881811 < −0.43790448 × (Function) + −0.04586086 × (Conditional, Loop, Debug, Simulation, Arrow, Hitting_wall) + 0.8505517 × (Big_cheese) + 0.097365424 × (Small_cheese)
head 3:	2.874901 < −0.017702527 × (Simulation, Hitting_wall) + 0.8470087 × (Big_cheese) + −0.4385394 × (Function) + 0.09731795 × (Small_cheese) + −0.06676157 × (Conditional) + −0.09061724 × (Loop) + −0.051380966 × (Debug) + −0.031886093 × (Arrow)

5. Discussion and Conclusions

This study presents a neural-symbolic AI (NSAI) approach for modeling learners' computational thinking knowledge, which learns from both symbolic educational knowledge and training data. It then compares the performance of the NSAI with deep neural networks trained only on training data, training data augmented with SMOTE upsampling, and an autoencoder.

Regarding the generalizability of the models, our findings indicate that the NSAI model, followed by the deep NN and deep NN-Autoencoder, exhibits the best generalizability among the considered models. The NSAI model demonstrates high accuracy and recall on unseen data, suggesting its potential to effectively predict both *high*- and *low*-performing learners. In the field of education, correctly identifying *low*-performing learners is crucial for providing timely interventions and support [1]. The high recall rates of the NSAI model make it a valuable tool in this regard, as it minimizes the risk of falsely identifying low performers. On the other hand, the other models show poor generalizability, particularly in identifying *low* performers. This indicates the limitations of using traditional deep learning models without additional techniques or knowledge incorporation to address the challenges posed by education data. This finding confirms the argument put forward by Venugopal et al. [83], Hooshyar and Yang [14], and Hooshyar [22], in that augmenting ANNs with symbolic knowledge can regularize them, improving their generalizability by achieving a higher accuracy and scalability by enabling them to learn from smaller datasets.

The superior generalizability of the NSAI model can be attributed to its incorporation of explicit knowledge in addition to the training data. By leveraging domain-specific knowledge, the NSAI model gains a deeper understanding of the underlying causal relationships among training examples. This broader understanding enables the model to generalize beyond the specific characteristics of the training data, making it more robust when faced with new or unseen data (e.g., [5]). In educational contexts, where datasets may be unrepresentative or exhibit different distributions, the NSAI model's ability to draw upon its explicit knowledge becomes particularly advantageous. For instance, a study conducted by Shakya et al. [69] proposes an innovative approach to enhance automated instruction

systems by accurately predicting student strategies and providing personalized support. Their neural-symbolic approach combines symbolic knowledge, using Markov models to represent the inherent relationships among input variables, with deep neural networks (specifically LSTMs). The integration of symbolic knowledge enables the model to capture complex relationships and patterns that may not be evident from the training data alone, resulting in improved accuracy and generalizability. These findings align with our research, highlighting the significance of incorporating domain knowledge into machine learning models, particularly neural networks, for educational applications. Regarding the impact of different data augmentation methods, while the autoencoder technique may slightly improve certain aspects of model performance on test data, the SMOTE method can cause a minor drop in performance compared to its performance on the original training data. This discrepancy can be attributed to a mismatch in the distribution of the training and test data. Data augmentation methods like SMOTE and autoencoders introduce synthetic or reconstructed samples to the training data. If the distribution of the augmented training data deviates from the distribution of the test data, the model may struggle to generalize well. Deep learning models are particularly sensitive to the distribution of the training data, and if the training data contain characteristics or patterns that are absent in the test data, the model's performance may suffer significantly. These findings align with the research of Ramezankhani et al. [84], which suggests that certain data augmentation methods can occasionally result in reduced model performance. Conversely, the NSAI approach appears to outperform all the trained models.

The analysis of data biases and interpretability provides insights into how the trained models rely on spurious correlations and ignore some essential features during their decision making. The correlation analysis reveals that certain features, such as Small_cheese, exhibit strong positive correlations with the class label (Final_score). Models that heavily rely on these strongly correlated features (especially the spurious correlations) during training may overlook the importance of other features and their relationships to the label. As underlined by Zhou et al. [85] and Hutt et al. [86], learning such biases can hamper models' generalizability to test data, as this causes the model to overlook crucial factors. As our findings show, the deep learning models that were only trained on these data not only mainly relied on the learned spurious correlations during their decision making, but also mostly ruled out many important features that are causally related to learner performance in CT. More explicitly, the models learn data-related biases during training and reflect these biases in their decision making. To address these biases and enhance interpretability, the NSAI model incorporates educational knowledge and adheres to educational restrictions during training. The representations and feature contributions learned by the model reflect its adherence to the underlying causal relationships between CT skills, concepts, and performance. The rules extracted from the NSAI model provide explicit guidelines for predicting high or low performance based on a combination of observable and unobservable features. These rules not only highlight the importance of CT skills and concepts but also shed light on the relevance of hidden heads that are related to learners' performance of other skills. In other words, the extracted rules from the NSAI approach can refine our initial educational knowledge as they provide weights for the rules, indicating that CT skills are more important in predicting learner performance compared to CT concepts.

The incorporation of educational knowledge into the NSAI model serves two key purposes. Firstly, it helps to alleviate the issue of data bias by providing a framework that ensures that the model considers a diverse range of factors beyond spurious correlated features. One of the main objectives of machine learning is to create reliable representations that accurately capture the causal relationship between input features and output labels. However, when models are trained on biased datasets, they may end up paying more attention to spurious correlations between input/output pairs that are not fundamentally relevant to the problem being solved [85]. The significance of using neural-symbolic AI for education lies in the potential to create more equitable and accurate systems that reflect

the true dynamics of learning processes, rather than perpetuating existing inequalities. The NSAI approach not only develops a more comprehensive understanding of learner performance by considering relevant factors but also addresses the class imbalance issue by learning the underlying relationships and patterns in the training data. This ensures a fair representation of all classes and improves the model's overall performance, which is crucial for educational tools that aim to effectively serve a diverse student population.

Secondly, the integration of educational knowledge into the model architecture enables the model to provide interpretable predictions. The model's decision-making process becomes transparent, as it can explain the learned representations that comply with the injected educational knowledge. As highlighted by several recent pieces of research (e.g., [1,14,23,24]), such interpretability is crucial in educational contexts, where stakeholders require explanations for model predictions to ensure transparency, fairness, and trustworthiness. This clarity is vital for fostering trust among educators and students, enabling them to rely on AI-driven insights for personalized learning experiences and interventions.

By combining the incorporation of knowledge with deep learning techniques, the NSAI model demonstrates how interpretability and generalizability can be improved in educational machine learning applications. The model's ability to capture complex relationships, consider multiple factors, and provide rule-based explanations facilitates a deeper understanding of learner performance. This not only aids in accurate predictions but also assists educators in identifying specific areas for intervention and support [23,87]. In conclusion, the NSAI model emerges as a promising approach, showcasing better generalizability and the ability to leverage domain-specific knowledge for trustworthy and interpretable predictions in educational contexts. The broader implication is the potential transformation of educational practices, making them more inclusive, personalized, and effective. Further research and development in this direction can contribute to the advancement of machine learning techniques in education, creating a more effective means of support for learners and educators.

Limitations and Future Works

One limitation of this study is its reliance on a specific dataset, which may affect the generalizability of the findings. Future research should address this by utilizing diverse educational datasets and examining various deep neural network models to enhance the robustness and applicability of the methods. Moreover, future work can focus on experimenting with different ways to inject educational knowledge into neural networks, potentially leveraging neural symbolic AI methods. Finally, incorporating the NSAI-driven learner modeling approach into digital learning platforms and assessing its real-world impact in classrooms is crucial. Thus, future work can evaluate the effectiveness of the NSAI approach in delivering personalized learning experiences that are unbiased, reliable, and easily understandable.

Author Contributions: Conceptualization, D.H., R.A. and Y.Y.; methodology, D.H. and Y.Y.; validation, D.H. and Y.Y.; formal analysis and data curation, D.H. and Y.Y.; writing—original draft preparation, D.H., R.A. and Y.Y.; writing—review and editing, D.H., R.A. and Y.Y.; supervision, D.H., R.A. and Y.Y.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Estonian Research Council grant (PRG2215).

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available due to privacy or ethical restrictions. The corresponding author can provide a sample of the dataset on reasonable request.

Acknowledgments: We would like to thank the Estonian Research Council for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vincent-Lancrin, S.; Van der Vlies, R. *Trustworthy Artificial Intelligence (AI) in Education: Promises and Challenges*; OECD: Paris, France, 2020.
2. Conati, C.; Lallé, S. 8. Student Modeling in Open-Ended Learning Environments. In *Handbook of Artificial Intelligence in Education*; Edward Elgar Publishing: Cheltenham, UK, 2023; pp. 170–183.
3. Abyaa, A.; Khalidi Idrissi, M.; Bennani, S. Learner Modelling: Systematic Review of the Literature from the Last 5 Years. *Educ. Technol. Res. Dev.* **2019**, *67*, 1105–1143. [CrossRef]
4. Azevedo, R.; Wiedbusch, M. Theories of Metacognition and Pedagogy Applied in AIED Systems. In *Handbook of Artificial Intelligence in Education*; Edward Elgar Publishing: Cheltenham, UK, 2023; pp. 45–67.
5. Garcez, A.d.A.; Bader, S.; Bowman, H.; Lamb, L.C.; de Penning, L.; Illuminoo, B.; Poon, H.; Zaverucha, C.G. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*; IOS Press: Amsterdam, The Netherlands, 2022; Volume 342, p. 327.
6. Lenat, D.B.; Prakash, M.; Shepherd, M. CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Mag.* **1985**, *6*, 65.
7. Bhanuse, R.; Mal, S. A Systematic Review: Deep Learning Based e-Learning Recommendation System. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 190–197.
8. Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity* **2019**, *2019*, 1306039. [CrossRef]
9. Hooshyar, D.; Huang, Y.-M.; Yang, Y. GameDKT: Deep Knowledge Tracing in Educational Games. *Expert Syst. Appl.* **2022**, *196*, 116670. [CrossRef]
10. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.J.; Sohl-Dickstein, J. Deep Knowledge Tracing. In Proceedings of the Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
11. Algarni, S.; Sheldon, F. Systematic Review of Recommendation Systems for Course Selection. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 560–596. [CrossRef]
12. Nielsen, M.A. *Neural Networks and Deep Learning*; Determination Press: San Francisco, CA, USA, 2015; Volume 25.
13. Wiedbusch, M.; Dever, D.; Li, S.; Amon, M.J.; Lajoie, S.; Azevedo, R. Measuring Multidimensional Facets of SRL Engagement with Multimodal Data. In *Unobtrusive Observations of Learning in Digital Environments: Examining Behavior, Cognition, Emotion, Metacognition and Social Processes Using Learning Analytics*; Springer: Cham, Switzerland, 2023; pp. 141–173.
14. Hooshyar, D.; Yang, Y. Neural-Symbolic Computing: A Step toward Interpretable AI in Education. *Bull. Tech. Comm. Learn. Technol.* **2021**, *21*, 2–6.
15. Torralba, A.; Efros, A.A. Unbiased Look at Dataset Bias. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1521–1528.
16. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4971–4980.
17. Gretton, A.; Györfi, L. Consistent Nonparametric Tests of Independence. *J. Mach. Learn. Res.* **2010**, *11*, 1391–1423.
18. Srivastava, M. Addressing Spurious Correlations in Machine Learning Models: A Comprehensive Review. *OSF Prepr.* **2023**. [CrossRef]
19. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLoS Med.* **2018**, *15*, e1002683. [CrossRef] [PubMed]
20. Ye, W.; Zheng, G.; Cao, X.; Ma, Y.; Hu, X.; Zhang, A. Spurious Correlations in Machine Learning: A Survey. *arXiv* **2024**, arXiv:2402.12715.
21. UNESCO. Beijing Consensus on Artificial Intelligence and Education. In Proceedings of the International Conference on Artificial Intelligence and Education, Planning Education in the AI Era: Lead the Leap, Beijing, China, 16–18 May 2019.
22. Hooshyar, D. Temporal Learner Modelling through Integration of Neural and Symbolic Architectures. *Educ. Inf. Technol.* **2024**, *29*, 1119–1146. [CrossRef]
23. Conati, C.; Porayska-Pomsta, K.; Mavrikis, M. AI in Education Needs Interpretable Machine Learning: Lessons from Open Learner Modelling. *arXiv* **2018**, arXiv:1807.00154.
24. Meltzer, J.P.; Tielemans, A. *The European Union AI Act: Next Steps and Issues for Building International Cooperation in AI*; Brookings Institution: Washington, DC, USA, 2022.
25. Rosé, C.P.; McLaughlin, E.A.; Liu, R.; Koedinger, K.R. Explanatory Learner Models: Why Machine Learning (Alone) Is Not the Answer. *Br. J. Educ. Technol.* **2019**, *50*, 2943–2958. [CrossRef]
26. Hooshyar, D.; Huang, Y.-M.; Yang, Y. A Three-Layered Student Learning Model for Prediction of Failure Risk in Online Learning. *Hum.-Centric Comput. Inf. Sci.* **2022**, *12*, 28.
27. Saarela, M.; Heilala, V.; Jääskelä, P.; Rantakaulio, A.; Kärkkäinen, T. Explainable Student Agency Analytics. *IEEE Access* **2021**, *9*, 137444–137459. [CrossRef]
28. Alwarthan, S.; Aslam, N.; Khan, I.U. An Explainable Model for Identifying At-Risk Student at Higher Education. *IEEE Access* **2022**, *10*, 107649–107668. [CrossRef]

29. Melo, E.; Silva, I.; Costa, D.G.; Viegas, C.M.D.; Barros, T.M. On the Use of Explainable Artificial Intelligence to Evaluate School Dropout. *Educ. Sci.* **2022**, *12*, 845. [CrossRef]
30. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–9 February 2020; pp. 180–186.
31. Lakkaraju, H.; Bastani, O. "How Do I Fool You?" Manipulating User Trust via Misleading Black Box Explanations. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–9 February 2020; pp. 79–85.
32. Hitzler, P.; Sarker, M.K. *Neuro-Symbolic Artificial Intelligence: The State of the Art*; IOS Press: Amsterdam, The Netherlands, 2022.
33. Garcez, A.d.A.; Lamb, L.C. Neurosymbolic AI: The 3rd Wave. *Artif. Intell. Rev.* **2023**, *56*, 12387–12406. [CrossRef]
34. Sarker, M.K.; Zhou, L.; Eberhart, A.; Hitzler, P. Neuro-Symbolic Artificial Intelligence. *AI Commun.* **2021**, *34*, 197–209. [CrossRef]
35. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
36. Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. *Found. Trends®Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
37. Hudon, A.; Phraxayavong, K.; Potvin, S.; Dumais, A. Comparing the Performance of Machine Learning Algorithms in the Automatic Classification of Psychotherapeutic Interactions in Avatar Therapy. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1119–1131. [CrossRef]
38. Werner, J.; Nowak, D.; Hunger, F.; Johnson, T.; Mark, A.; Gösta, A.; Edelvik, F. Predicting Wind Comfort in an Urban Area: A Comparison of a Regression-with a Classification-CNN for General Wind Rose Statistics. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 98–125. [CrossRef]
39. Mehta, D.; Klarmann, N. Autoencoder-Based Visual Anomaly Localization for Manufacturing Quality Control. *Mach. Learn. Knowl. Extr.* **2023**, *6*, 1–17. [CrossRef]
40. Rodriguez-Bazan, H.; Sidorov, G.; Escamilla-Ambrosio, P.J. Android Malware Classification Based on Fuzzy Hashing Visualization. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1826–1847. [CrossRef]
41. Samkari, E.; Arif, M.; Alghamdi, M.; Al Ghamdi, M.A. Human Pose Estimation Using Deep Learning: A Systematic Literature Review. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1612–1659. [CrossRef]
42. Zhou, Q.; Wang, J.; Yu, X.; Wang, S.; Zhang, Y. A Survey of Deep Learning for Alzheimer's Disease. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 611–668. [CrossRef]
43. Mapundu, M.T.; Kabudula, C.W.; Musenge, E.; Olago, V.; Celik, T. Explainable Stacked Ensemble Deep Learning (SEDL) Framework to Determine Cause of Death from Verbal Autopsies. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1570–1588. [CrossRef]
44. Nemirovsky-Rotman, S.; Bercovich, E. Explicit Physics-Informed Deep Learning for Computer-Aided Diagnostic Tasks in Medical Imaging. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 385–401. [CrossRef]
45. Bendangnuksung, P.P. Students' Performance Prediction Using Deep Neural Network. *Int. J. Appl. Eng. Res.* **2018**, *13*, 1171–1176.
46. Wang, L.; Sy, A.; Liu, L.; Piech, C. Deep Knowledge Tracing on Programming Exercises. In Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, Cambridge, MA, USA, 20–21 April 2017; pp. 201–204.
47. Kukkar, A.; Mohana, R.; Sharma, A.; Nayyar, A. Prediction of Student Academic Performance Based on Their Emotional Wellbeing and Interaction on Various E-Learning Platforms. *Educ. Inf. Technol.* **2023**, *28*, 9655–9684. [CrossRef]
48. Fei, M.; Yeung, D.-Y. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; pp. 256–263.
49. Teruel, M.; Alonso Alemany, L. Co-Embeddings for Student Modeling in Virtual Learning Environments. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore, 8–11 July 2018; pp. 73–80.
50. Whitehill, J.; Mohan, K.; Seaton, D.; Rosen, Y.; Tingley, D. Delving Deeper into MOOC Student Dropout Prediction. *arXiv* **2017**, arXiv:1702.06404.
51. Abhinav, K.; Subramanian, V.; Dubey, A.; Bhat, P.; Venkat, A.D. LeCoRe: A Framework for Modeling Learner's Preference. In Proceedings of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA, 15–18 July 2018.
52. Wong, C. Sequence Based Course Recommender for Personalized Curriculum Planning. In *Artificial Intelligence in Education. AIED 2018, Proceedings of the AIED 2018, London, UK, 27–30 June 2018*; Springer: Cham, Switzerland, 2018; pp. 531–534.
53. Hooshyar, D.; Yang, Y. ImageLM: Interpretable Image-Based Learner Modelling for Classifying Learners' Computational Thinking. *Expert Syst. Appl.* **2024**, *238*, 122283. [CrossRef]
54. Taghipour, K.; Ng, H.T. A Neural Approach to Automated Essay Scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1882–1891.
55. Zhao, S.; Zhang, Y.; Xiong, X.; Botelho, A.; Heffernan, N. A Memory-Augmented Neural Model for Automated Grading. In Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, Cambridge, MA, USA, 20–21 April 2017; pp. 189–192.
56. Mao, Y. Deep Learning vs. Bayesian Knowledge Tracing: Student Models for Interventions. *J. Educ. Data Min.* **2018**, *10*, 28–54.
57. Tang, S.; Peterson, J.C.; Pardos, Z.A. Deep Neural Networks and How They Apply to Sequential Education Data. In Proceedings of the Third (2016) ACM Conference on Learning @ Scale, Edinburgh, UK, 25–26 April 2016; pp. 321–324.
58. Sharada, N.; Shashi, M.; Xiong, X. Modeling Student Knowledge Retention Using Deep Learning and Random Forests. *J. Eng. Appl. Sci.* **2018**, *13*, 1347–1353.

59. Alam, M.M.; Mohiuddin, K.; Das, A.K.; Islam, M.K.; Kaonain, M.S.; Ali, M.H. A Reduced Feature Based Neural Network Approach to Classify the Category of Students. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, Shanghai, China, 9–12 March 2018; pp. 28–32.
60. Kim, B.-H.; Vizitei, E.; Ganapathi, V. GritNet: Student Performance Prediction with Deep Learning. *arXiv* **2018**, arXiv:1804.07405.
61. Wang, W.; Yu, H.; Miao, C. Deep Model for Dropout Prediction in MOOCs. In Proceedings of the 2nd International Conference on Crowd Science and Engineering, Beijing, China, 6–9 July 2017; pp. 26–32.
62. Tato, A.; Nkambou, R.; Dufresne, A.; Beauchamp, M.H. Convolutional Neural Network for Automatic Detection of Sociomoral Reasoning Level. In Proceedings of the 10th International Conference on Educational Data Mining, Wuhan, China, 25–28 June 2017.
63. Min, W.; Wiggins, J.B.; Pezzullo, L.G.; Vail, A.K.; Boyer, K.E.; Mott, B.W.; Frankosky, M.H.; Wiebe, E.N.; Lester, J.C. Predicting Dialogue Acts for Intelligent Virtual Agents with Multimodal Student Interaction Data. In Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, 29 June–2 July 2016.
64. Sharma, A.; Biswas, A.; Gandhi, A.; Patil, S.; Deshmukh, O. LIVELINET: A Multimodal Deep Recurrent Neural Network to Predict Liveliness in Educational Videos. In Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, 29 June–2 July 2016.
65. Tran, S.N.; Garcez, A.S.d.A. Deep Logic Networks: Inserting and Extracting Knowledge from Deep Belief Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *29*, 246–258. [CrossRef]
66. Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; Xing, E. Harnessing Deep Neural Networks with Logic Rules. *arXiv* **2016**, arXiv:1603.06318.
67. Serafini, L.; d'Avila Garcez, A.S. Learning and Reasoning with Logic Tensor Networks. In *AI*IA 2016 Advances in Artificial Intelligence. AI*IA 2016, Proceedings of the XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, 29 November–1 December 2016*; Adorni, G., Cagnoni, S., Gori, M., Maratea, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 334–348.
68. Yu, D.; Yang, B.; Liu, D.; Wang, H.; Pan, S. A Survey on Neural-Symbolic Learning Systems. *Neural Netw.* **2023**, *166*, 105–126. [CrossRef] [PubMed]
69. Shakya, A.; Rus, V.; Venugopal, D. Student Strategy Prediction Using a Neuro-Symbolic Approach. In Proceedings of the 14th International Conference on Educational Data Mining, Online, 29 June–2 July 2021.
70. Hooshyar, D.; Lim, H.; Pedaste, M.; Yang, K.; Fathi, M.; Yang, Y. AutoThinking: An Adaptive Computational Thinking Game. In *Innovative Technologies and Learning. ICITL 2019, Proceedings of the Second International Conference, ICITL 2019, Tromsø, Norway, 2–5 December 2019*; Rønningsbakk, L., Wu, T.T., Sandnes, F., Huang, Y.M., Eds.; Springer: Cham, Switzerland, 2019; pp. 381–391.
71. Towell, G.G.; Shavlik, J.W. Knowledge-Based Artificial Neural Networks. *Artif. Intell.* **1994**, *70*, 119–165. [CrossRef]
72. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-Agnostic Interpretability of Machine Learning. *arXiv* **2016**, arXiv:1606.05386.
73. Azevedo, R.; Dever, D. Metacognition in Multimedia Learning. In *Cambridge Handbook of Multimedia*; Cambridge University Press: Cambridge, UK, 2022; pp. 132–141.
74. Greene, J.A.; Azevedo, R. A Macro-Level Analysis of SRL Processes and Their Relations to the Acquisition of a Sophisticated Mental Model of a Complex System. *Contemp. Educ. Psychol.* **2009**, *34*, 18–29. [CrossRef]
75. Pekrun, R. The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educ. Psychol. Rev.* **2006**, *18*, 315–341. [CrossRef]
76. Winne, P.H.; Azevedo, R. Metacognition. In *Cambridge Handbook of the Learning Sciences*; Cambridge University Press: Cambridge, UK, 2014; pp. 63–87.
77. Hooshyar, D. Effects of Technology-enhanced Learning Approaches on Learners with Different Prior Learning Attitudes and Knowledge in Computational Thinking. *Comput. Appl. Eng. Educ.* **2022**, *30*, 64–76. [CrossRef]
78. European Union Agency for Fundamental Rights. *Bias in Algorithms—Artificial Intelligence and Discrimination*; European Union Agency for Fundamental Rights: Vienna, Austria, 2022.
79. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (Technology) Is Power: A Critical Survey of "Bias" in Nlp. *arXiv* **2020**, arXiv:2005.14050.
80. Johnson, J.M.; Khoshgoftaar, T.M. Survey on Deep Learning with Class Imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
81. Liusie, A.; Raina, V.; Raina, V.; Gales, M. Analyzing Biases to Spurious Correlations in Text Classification Tasks. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Online, 20–23 November 2022; pp. 78–84.
82. Hooshyar, D.; El Mawas, N.; Milrad, M.; Yang, Y. Modeling Learners to Early Predict Their Performance in Educational Computer Games. *IEEE Access* **2023**, *11*, 20399–20417. [CrossRef]
83. Venugopal, D.; Rus, V.; Shakya, A. Neuro-Symbolic Models: A Scalable, Explainable Framework for Strategy Discovery from Big Edu-Data. In Proceedings of the 2nd Learner Data Institute Workshop in Conjunction with the 14th International Educational Data Mining Conference, Online, 29 June–2 July 2021.
84. Ramezankhani, A.; Pournik, O.; Shahrabi, J.; Azizi, F.; Hadaegh, F.; Khalili, D. The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes. *Med. Decis. Mak.* **2016**, *36*, 137–144. [CrossRef]
85. Zhou, C.; Ma, X.; Michel, P.; Neubig, G. Examining and Combating Spurious Features under Distribution Shift. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; pp. 12857–12867.

86. Hutt, S.; Gardner, M.; Duckworth, A.L.; D'Mello, S.K. Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. In Proceedings of the 12th International Conference on Educational Data Mining, Montreal, QC, Canda, 2–5 July 2019.
87. Fiok, K.; Farahani, F.V.; Karwowski, W.; Ahram, T. Explainable Artificial Intelligence for Education and Training. *J. Def. Model. Simul.* **2022**, *19*, 133–144. [CrossRef]