Journal of Integrative Neuroscience

J. Integr. Neurosci. **2026**; 25(9): 52375
https://doi.org/10.31083/JIN52375

Original Research

Askisi-MD: Development and Validation of a Web-Based Neuropsychological Screener for Mathematical Learning Difficulties Using Explainable Machine Learning

Nikolaos C. Zygouris[1,*], Stefanos K. Styliaras[1], Filippos Vlachos[2], Evaggelos Spyrou[3], Panayiotis Patrikelis[4], Lambros Messinis[4], Grigorios Nasios[5]

[1]Laboratory of Digital Neuropsychological Assessment, Department of Informatics and Telecommunications, University of Thessaly, 35100 Lamia, Greece

[2]Department of Special Education, University of Thessaly, Argonafton & Filellinon, 38221 Volos, Greece

[3]Department of Informatics and Telecommunications, University of Thessaly, 35100 Lamia, Greece

[4]Laboratory of Neuropsychology and Behavioral Neuroscience, School of Psychology, Aristotle University of Thessaloniki, 57001 Thessaloniki, Greece

[5]Department of Speech and Language Therapy, University of Ioannina, 45500 Ioannina, Greece

*Correspondence: nzygouris@uth.gr (Nikolaos C. Zygouris)

Academic Editor: Bettina Platt

Submitted: 29 March 2026 Revised: 29 June 2026 Accepted: 10 July 2026 Published: 3 September 2026

Abstract

Background: Mathematical learning difficulties (MLD) are associated with weaknesses in both domain-specific numerical abilities and domain-general neurocognitive processes. In this study, we developed and validated a web-based neuropsychological screener, Askisi-Mathematical Difficulties (MD), that integrates cognitive control and mathematics-related tasks for the first-line identification of children at risk for MLD was developed and validated. **Methods:** We enrolled 564 children, including 282 children with an established MLD diagnosis recruited from a state diagnostic center and 282 typically developing children pair-matched for age and gender. Askisi-MD consisted of six tasks, including three cognitive tasks assessing inhibitory control, visual pattern recognition, and visual working memory, as well as three mathematics-related tasks assessing mental calculation verification, mathematical terminology comprehension, and arithmetic word problem operation selection. Accuracy scores and latency indicators were analyzed. Differences among groups were examined using Welch's independent-samples *t*-tests, Benjamini–Hochberg False Discovery Rate (FDR) correction, and Hedges' g effect sizes. Internal coherence was examined using McDonald's Ω. Discriminant validity was evaluated via receiver operating characteristic (ROC) analysis and through the use of classification indices. Supervised machine learning models, including Support Vector Classifier, Random Forest, and eXtreme Gradient Boosting (XGBoost) classifiers, were trained and cross-validated, while model interpretability was examined via SHapley Additive exPlanations (SHAP) analysis. **Results:** Children with MLD showed significantly poorer performance across all six task-score indicators after FDR correction, with moderate-to-large effect sizes. Minute-based task-completion time differences were more selective. Children with MLD showed significantly longer completion times on the Mental Calculation and Arithmetic Word Problem tasks, whereas completion-time differences were not significant for Visual Pattern Recognition Abilities or Mathematical Terminology Comprehension. McDonald's Ω indicated modest internal coherence for the neuropsychological task-score composite and the total task-score composite. The strongest internal coherence was observed for the mathematical processing-efficiency composite. ROC analysis supported strong case-control discrimination, and classification indices indicated balanced group classification. Across the cross-validation folds, XGBoost achieved the highest mean accuracy, precision, and F1 score, while XGBoost and Random Forest demonstrated the same mean recall. XGBoost also showed strong discriminative performance, as indicated by the cross-validated area under the receiver operating characteristic curve (ROC-AUC). SHAP analysis indicated that Arithmetic Word-Problem Task Latency, Arithmetic Word Problem performance, Mathematical Terminology Comprehension, Mental Calculation, and Go/No-Go error score had greater influence on XGBoost model predictions than the remaining task indicators. **Conclusions:** Askisi-MD provides a neuropsychologically informed web-based approach to MLD screening by jointly sampling mathematical performance and supporting cognitive control functions. The composite-level internal coherence findings support the use of Askisi-MD as a multidomain screener rather than as a set of independent mathematics subscales. Its explainable machine learning framework further enhances model interpretability by identifying the task-level indicators most strongly contributing to classification outcomes. Askisi-MD thus offers promise as a first-line screening tool to support referral decisions, rather than a stand-alone diagnostic instrument.

Keywords: mathematical learning difficulties; learning disabilities; neuropsychological tests; machine learning; early screening

1. Introduction

Mathematical learning difficulties (MLD) refer to persistent difficulties in acquiring numerical and arithmetic skills that are not attributable to general intellectual disabilities, poor education, or sensory deficits. These difficulties may manifest as challenges in understanding number magnitude, performing basic calculations, or recalling arithmetic facts, and often appear early in childhood [1]. Children affected by MLD exhibit impairments in core numerical processing abilities, such as the ability to rapidly quantify small sets, number comparison, and symbolic number understanding, all of which have the potential to significantly impact their academic trajectory and self-esteem. Effectively identifying MLD and intervening at an early stage is therefore crucial to mitigate any long-term educational and social consequences [2]. MLD affects an estimated 3% to 7% of school-age children, although its prevalence varies somewhat based on diagnostic criteria, assessment tools, and population samples [3]. While mathematical difficulties have been reported to be as common as reading disabilities, they remain significantly underdiagnosed and under-researched [4,5,6].

Luciano et al. (2025) [7] emphasized that children with MLD frequently present with co-occurring learning difficulties, including dyslexia, and comorbid neurodevelopmental disorders such as attention-deficit/hyperactivity disorder, complicating accurate assessment and intervention efforts. MLD underdiagnosis may be more common among females due to gender biases in teacher referrals and differences in coping strategies [8]. Importantly, mathematical difficulties are not solely an academic issue, as they are also linked to long-term outcomes, including lower educational attainment, reduced employment prospects, and financial difficulties in adulthood [3,6,7,9]. Given how often MLD presents contemporaneously with other neurodevelopmental disorders, large-scale screening and improved epidemiological tracking are essential [6].

Several studies have asserted the importance of shifting from the relatively one-dimensional "dyscalculia" term in favor of the broader multidimensional MLD construct to more effectively capture the heterogeneous pathways through which children present difficulties in mathematics [10,11,12,13]. In line with this perspective, MLD encompasses deficits in calculation but may also involve disorders across multiple domains of mathematical ability, including number processing, procedural knowledge, visuospatial organization, memory for arithmetic facts, and higher-order mathematical reasoning. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [14] describes dyscalculia under specific learning disorders with impairment in mathematics, characterized by persistent and clinically significant difficulties in number sense, memorization of arithmetic facts, accurate or fluent calculation, and accurate mathematical reasoning [14]. In the present study, the term MLD is used preferentially given that it more effectively describes the broader educational and neurocognitive profile of children who experience significant challenges in mathematics.

Two of the most well-studied neuropsychological models of numerical cognition suggest that number processing relies on distinct representational and format-specific systems. These neural networks exist across specialized regions in the left and right cerebral hemispheres [15,16]. These models are grounded primarily in findings from adult patients who have suffered traumatic brain injuries and are based on observed dissociations in number-related functions such as calculation, numerical comprehension, or recognition of digits. These patterns support the notion that different aspects of numerical processing are localized to separate neural circuits and can be selectively disrupted by localized brain injury.

The McCloskey, Caramazza, and Basili (1985) [17] model offers a neuropsychological account of how numerical information is processed in the brain. It proposes that number processing involves distinct cognitive systems, including one responsible for number comprehension, one involved in number production, and another that is required to perform calculations. These systems operate on abstract numerical representations that are separate from the physical forms in which numbers appear, be they spoken, written, or symbolic. This separation is important because deficits in specific brain regions can selectively impair one system while leaving others relatively intact. According to this model, lesions in left parietal areas may affect calculation abilities without disrupting number comprehension or production. Affected individuals may retain the ability to recognize and read numbers but fail to perform even simple arithmetic operations. This model has become important for understanding numerical deficits in individuals with acquired brain injuries and developmental disorders like MLD, emphasizing that distinct neural circuits highlight different aspects of number processing [17].

The Dehaene model of numerical cognition offers a neuropsychological explanation for how numbers are understood and processed by the neurocognitive system. This so-called "Triple Code Model" suggests that the brain represents numbers in three main ways, each of which is linked to different cognitive functions and brain regions. The first of these is the analog magnitude representation, which encompasses a nonverbal sense of quantity that does not require the need to count. This sense is believed to be rooted in the intraparietal sulcus (IPS) and forms what is often described as a "mental number line". The second is the visual Arabic code, which involves recognizing written numerals and is linked to the occipitotemporal cortex, assisting in reading and distinguishing numbers on a page. The third is the verbal code, which is associated with the auditory and phonological representation of numbers and plays a central role in tasks such as counting. This process relies on language areas in the left perisylvian cortex. Together, these three systems operate as an integrated neural network that supports the comprehension, mental manipulation, and expression of numerical information. When one of these systems is impaired, it can lead to difficulties such as dyscalculia [18,19,20].

Research focused on children has suggested that numerical disabilities are linked to clear differences in the functionality of the brain's number networks. Neuroimaging studies have demonstrated that the parietal cortex, specifically the IPS, serves as a hub for quantity processing [18]. In children with numerical disabilities, this region often responds less efficiently than in normal controls when comparing magnitudes. The same impaired activation is present when they perform simple calculations, and its structure can also be subtly altered [19]. This is consistent with both the McCloskey model and Dehaene's Triple Code Model, which posit that parietal circuits play a crucial role in representing abstract numerical magnitude and supporting calculation.

Numerical difficulties are not confined to a single brain region. Functional MRI and connectivity studies have shown that children with numerical disabilities often exhibit disrupted communication within a broader fronto-parietal network [21]. Atypical patterns have been reported between the IPS and frontal regions involved in attention, working memory, and strategy control, as well as temporal and occipitotemporal areas that support symbolic number forms [22,23,24]. Some children show signs of compensatory neural activation, recruiting additional frontal or temporal regions to cope with tasks that are handled more automatically by parietal circuits in typically developing peers [25]. Others present with hyperconnectivity or hyperactivity in regions like the hippocampus, suggesting a greater reliance on strategies that depend on effortful memory processes rather than efficient quantity processing in these individuals [26,27].

Longitudinal and training studies have emphasized neuroplasticity in the management of MLD, suggesting that the observed differences in brain function may attenuate or shift with instruction and practice [28]. Children with numerical disabilities who receive targeted numerical interventions, such as number-line or arithmetic training, may exhibit neural network reorganization in the form of increased temporoparietal activation, changes in structural covariance, and/or reductions in abnormal hyperconnectivity after successful training [24,27,29]. The studies suggest that numerical disabilities in children can best be understood as a consequence of the atypical development and coordination of the neural circuits detailed in these neuropsychological models, rather than a simple lack of practice with numbers.

In addition to mathematical challenges, children with MLD face broader difficulties in how they process and manage information. Neuropsychological research has consistently found that a large subgroup of children with numerical disabilities have problems with working memory, attention, and executive function [30,31]. Owing to these difficulties, even relatively simple classroom tasks, like keeping track of the steps in a subtraction problem or following multi-step instructions, can become unexpectedly demanding. Issues with working memory are among the most consistent findings in these children, especially when they need to remember both items and their order. Children with MLD often exhibit weaker performance on tasks such as recalling a sequence of digits or reproducing a pattern of blocks, particularly when order is important [32,33]. Impacts on spatial working memory, which is important for tasks such as lining numbers up correctly, reading lines of numbers, or understanding place value, are also frequently reported [34]. Recent work in different cultural contexts, including Egyptian and Moroccan samples, has provided support for the notion that reduced working memory capacity is a stable feature in many children with numerical disabilities, rather than something that is simply attributable to insufficient schooling or motivational factors [35].

Inhibition, cognitive flexibility, and updating also appear to be compromised in many children with MLD. Indeed, MLD is commonly associated with greater difficulty inhibiting irrelevant responses, shifting between tasks, and managing complex problem-solving demands compared with their typically developing peers [36,37,38]. Longitudinal and screening studies have suggested that weaknesses in these number-related executive functions can help differentiate between cchildren meeting the study's MLD criteria and those whose poor mathematics performance reflects factors other than specific learning difficulties [39,40].

It is important to note that there is no single cognitive profile associated with MLD. Some children show relative deficits only in number processing, whereas their general cognition is intact, while others exhibit a broader pattern of impairment that entails language, attention, or global executive dysfunction [30,37,41]. These cognitive weaknesses also frequently interact with emotional factors such as anxiety. When a child's working memory resources are already limited, the stress and worry that come with repeated experiences of failure in mathematics can further disrupt this system, leaving even fewer cognitive resources available for mathematical problem solving. Children with MLD often face a dual challenge stemming from the association between math-related anxiety and working memory. Specifically, it has been proposed that these children begin with reduced working memory resources and experience greater anxiety in math-related situations, which can further strain their limited capacity for problem solving [42]. Over time, this combination can reinforce math avoidance, further widening the gap in their abilities relative to their peers [43].

Over the past two decades, there has been a clear shift from traditional pencil-and-paper testing to computer-based platforms when seeking to identify children at risk for

MLD. One of the earliest and best-studied software applications is the Dyscalculia Screener, which is a standardized computer-based assessment for children approximately 6–14 years of age that was originally developed by Butterworth (2003) [44]. It uses timed tasks focused on core number processing skills and arithmetic fluency to provide risk classifications rather than a full curriculum-based math profile [45,46,47]. Building on this approach, Räsänen et al. [48] have developed a large-scale online dyscalculia screener, delivered entirely via the internet, which can be administered to whole cohorts from Grade 3 to Grade 9 and has been psychometrically analyzed to reveal separable factors for basic number processing and arithmetic [49].

More recently, game-based and immersive digital environments have been proposed to increase engagement and make screening more acceptable in school settings. For instance, a web-based educational game designed specifically for the early screening and support of dyscalculia has been developed in which numerical tasks are embedded in an interactive narrative, while performance data are collected in the background [50]. Separately, a 3D dyscalculia identification game framework has been described that integrates screening capabilities into an immersive environment with a focus on counting, comparison, and simple arithmetic in spatially rich situations [51]. Other systems use mobile or multi-platform solutions, such as a mobile-based screening and refinement system that delivers mini-games on smartphones or tablets to estimate the risk of dyscalculia and dysgraphia [52] or a computer game-based solution for the early diagnosis of learning disabilities, including dyscalculia, based on performance patterns across several game tasks, as described by Samarasinghe and Abeyasinghe [53].

In addition to these screeners, a second group of computer-based tools has been developed primarily for the purposes of intervention while also incorporating diagnostically informative features. Two of the best-studied of these tools include The Number Race and Calcularis/Rescue Calcularis, both of which are adaptive computer programs that train core numerical skills in children with dyscalculia or marked mathematical difficulties [54,55]. These computer programs estimate a learner model of numerical knowledge and employ dynamic adjustment of task difficulty. Additionally, they generate detailed profiles of strengths and weaknesses such that they function as both training tools and computer-based assessments of number processing. Additional platforms such as Dots2Digit and Dots2Track provide web-based exercises in basic numerosity and early arithmetic and are often used with children at risk of dyscalculia. Although not formal diagnostic instruments, they contribute performance data that can complement dedicated screeners [48,56,57,58,59]. Taken together, these tools illustrate a variety of computer-based screeners and adaptive game environments that leverage the advantages of digital delivery for the early identification and support of children with MLD. Although existing screeners and gamified tools demonstrate that brief digital tasks can identify children at risk, the few applications developed for Greek-language educational settings have primarily focused on spelling and reading difficulties rather than MLD [58,59]. Askisi-Mathematical Difficulties (MD) is grounded in a hybrid neuropsychological framework, as it samples mathematical reasoning and math-language decisions while also capturing executive and visuospatial deficits that often shape error patterns, fatigue, and time-cost in real classrooms.

Study Objectives and Hypothesis

The primary aim of this study was to develop and validate Askisi-MD, a web-based neuropsychological screening application designed to identify school-age children with MLD. Askisi-MD is designed to integrate brief measures of mathematical performance and domain-relevant cognitive functions that are frequently implicated in MLD. Unlike existing instruments that focus primarily on mathematical attainment, Askisi-MD adopts a hybrid screening framework that combines conventional mathematics-related tasks with cognitive indicators that may contribute to persistent learning problems. The screening battery was developed to sample multiple domains within a brief administration time, to reduce test burden and fatigue. Askisi-MD was also designed to fill a practical neuropsychological gap, as current screening strategies typically rely either on achievement-heavy mathematics assessment or time-intensive clinical assessment, whereas there remains a lack of screening tools that integrate cognitive mechanisms known to constrain mathematical learning alongside curriculum-grounded mathematics decisions in a single, scalable session.

Given the availability of validated Greek-language tools for MLD screening, another objective of this study was to characterize neuropsychological differences between typically developing children and children with established MLD in a Greek-speaking population. These group-level profiles can be used to inform and substantiate the construction of a screening approach grounded in neurocognitive theory. The web-based format enables broad accessibility, standardized delivery, and scalable use across devices as it only requires access to a web browser, facilitating dissemination without the constraints of platform-specific installation [58,59].

To evaluate predictive validity and to move beyond traditional score threshold-based approaches, this study additionally employed three supervised machine learning classifiers trained on the task-derived features, including Support Vector Classifier (SVC), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) models. SVC models map input features into a high-dimensional space to find an optimal separating hyperplane. RF models employ an ensemble of decision trees built via bootstrap aggregation, providing a nonlinear baseline that can capture interactions among cognitive and mathematical indicators. XGBoost models are high-performance models that often yield improved discrimination on structured behavioral data. Comparisons of these models were used to more systematically determine whether Askisi-MD supports accurate classification under different learning assumptions while providing stronger evidence based on the fact that discrimination is not dependent on a single analytic approach.

This study was framed around three hypotheses. First, we hypothesized that Greek children with an established MLD diagnosis derived from written testing at a state diagnostic center would exhibit poorer performance (fewer correct responses/lower scores) and longer minute-based task-completion time on Askisi-MD tasks relative to typically developing controls. Second, we hypothesized that task-derived features from Askisi-MD would be sufficient to enable reliable discrimination between children with MLD and controls. Specifically, we posited that classification models would be able to achieve good overall discrimination, as indexed by ROC analysis, while showing strong threshold-based performance, correctly assigning children with MLD and controls to appropriate groups at rates meaningfully above chance. We further expected these models to capture multivariate and potentially nonlinear relationships between the cognitive–mathematical features and group membership. Consistent performance across all three models would further support the robustness of Askisi-MD as a screening tool.

2. Materials and Methods

2.1 Participants

A total of 564 right-handed children (287 male, 277 female) from 9–12 years of age (mean: 10.19 years, SD: 1.15 years) participated in the study. The experimental group consisted of 282 children with an established MLD diagnosis from a state diagnostic center (in accordance with Greek Laws 4547/2018 and 4823/2021).The control group comprised 282 children without diagnosed learning difficulties or other neurodevelopmental disorders who were randomly recruited from the same schools as the children with MLD. Controls were frequency-matched to the MLD group at the group level by age and gender. Controls were selected within the same age and gender categories represented in the MLD group, without tolerance outside these categories and without replacement. No individual case–control pairs were created or retained. Therefore, the groups were treated as independent samples in all statistical analyses. All participants attended regular school placements and had no documented medical or psychiatric conditions according to their school medical records. Participation in the study required informed consent from all participants' parents/guardians, in line with ethical standards.

Comorbid neurodevelopmental conditions are common in referral populations and can systematically influence executive task performance. To reduce this source of heterogeneity and to avoid attributing broader attentional/reading-related variance to mathematics-specific impairment, recruitment for the MLD group was conducted through a state diagnostic center. The selection of students was based on records from the Centers for Interdisciplinary Assessment, Counseling, and Support (KE.DA.SY.), and all necessary approvals were obtained before selection.

In accordance with Greek law, KE.DA.SY. uses interdisciplinary teams to conduct individual assessments of preschool and school-age students. These teams consist of a special education teacher, a psychologist, and a social worker. The special education teacher is responsible for evaluating reading, writing, spelling, and mathematical performance using both standardized and non-standardized tests, while the psychologist conducts cognitive and behavioral assessments using batteries and tests such as the Greek version of the Wechsler Intelligence Scale for Children (WISC-III) [60] and the Greek version of attention-deficit/hyperactivity disorder (ADHD) Rating Scale-IV [61,62], which are completed by the children's parents and teachers. Recruitment and data collection were conducted between 2020 and 2025. Children in the MLD group had received a diagnosis from an official Greek public diagnostic service within the two to three years preceding their participation. These diagnoses were not independently reconfirmed by the research team. Group allocation was based on a review of the available multidisciplinary diagnostic documentation. Although participants in the MLD group had been previously identified through official diagnostic procedures, it should be noted that their clinical records included WISC-III-based information and discrepancy-oriented interpretations. The research team did not administer the WISC-III or apply an intelligence quotient (IQ)–achievement discrepancy criterion for participant classification. Instead, group allocation was based solely on existing multidisciplinary diagnostic documentation from public diagnostic services. Importantly, the detailed multidisciplinary reports did not rely exclusively on an IQ–achievement discrepancy interpretation. Instead, they also included functional evidence consistent with guidance from the DSM-5 [14], such as persistent mathematical difficulties, school-based impairment, developmental and educational information, behavioral observations, and exclusionary considerations recorded by the multidisciplinary team. MLD classification in this study was therefore based on the official diagnostic decision of the Greek public service and on the broader functional profile documented in each child's file, rather than on the WISC-III discrepancy criterion alone. Children who were not native Greek speakers or lived in bilingual/multilingual family environments were not included in the study cohort. Participants were also excluded if they had any other comorbid developmental or behavioral disorders.

Fig. 1. A representative mental calculation task. The child has to choose whether the result was correct (left button) or incorrect (right button).

This approach enabled the more controlled clinical characterization of MLD and reduced the likelihood that the case definition would be driven primarily by non-mathematical neurodevelopmental factors. In the present study, MLD was considered to be present if children had received a formal written determination of significant and persistent mathematical learning difficulties from the center prior to study recruitment. This observational study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (see **Supplementary Table 1**).

2.2 Askisi-MD Neuropsychological Web-Based Screener

The Askisi-MD screener was used to evaluate the numerical and cognitive skills of participating children using a web-based neuropsychological battery of six tasks. These tasks were designed to evaluate essential skills related to mathematical difficulties. To familiarize participants with the testing procedure, children first completed a brief training trial in which they were asked to click on a picture. This trial also served to verify basic proficiency with mouse use. If a child was unable to correctly use the mouse to click on at least three of the six picture items, the screening session was not initiated. It should be noted that no children were excluded from the testing procedure, as all participants had prior experience using computers and were able to complete the digital tasks.

During the main screening test procedure, children's correct responses were measured across six tasks focused on cognitive (the first three presented tasks) and numerical (the last three presented tasks) performance. These tasks were as follows:

(1) A cognitive control and response selection (Go/No-Go) task. Children were instructed to respond only when a target stimulus (a book) appeared ("Go") and to withhold responding when any non-target image appeared ("No-Go"; ruler, pencil, globe, pen, eraser). Stimuli were presented in a random order. Performance on the Go/No-Go task was scored as an error index. The total score reflected incorrect responses across the 10 Go trials and 20 No-Go trials, including omission errors on Go trials and commission errors on No-Go trials. Scores could range from 0 to 30, with higher scores indicating poorer inhibitory-control performance and lower scores indicating better response selection and inhibition. Therefore, this variable was reverse-directional relative to the other task-score indicators, as higher Go/No-Go values reflected poorer performance. Response latency was not recorded in this version of the platform.

(2) A Visual Pattern Recognition Ability task. Children were asked to complete a series of visual diagrams (5 items) and patterns (6 items) by choosing the appropriate missing component from six alternative options presented for each trial.

(3) A Visual Working Memory Task. Children were presented with 22 numerical sequences that increased in length from three to eight digits. After each sequence, they reproduced the exact sequence using a numerical keypad (0–9). The task was terminated if a child made two consecutive errors or three errors in total, in order to reduce fatigue or discomfort and to maintain a consistent testing procedure across participants.

(4) A Mental Calculation task. Children were presented with 4 arithmetic calculations and were required to judge whether a presented solution to an arithmetic operation was correct or incorrect (Fig. 1)

(5) A Mathematical Terminology Comprehension task. Children were presented with two questions with four possible answers and had to correctly define numerical concepts (Fig. 2).

(6) An Arithmetic Word Problem task. Two mathematical problems were presented, and children had to decide whether addition, subtraction, multiplication, or division was needed to obtain the solution (Fig. 3).

The mathematical tasks were purposely brief (2–4 items) to reduce burden and maintain engagement in a school-feasible workflow. These components were designed as focused "micro-screeners" focused on specific classroom-relevant decisions, including mathematical language, operation selection, and calculation verification, rather than as long-form achievement or mastery tests. Therefore, these task-level scores were interpreted as indi-

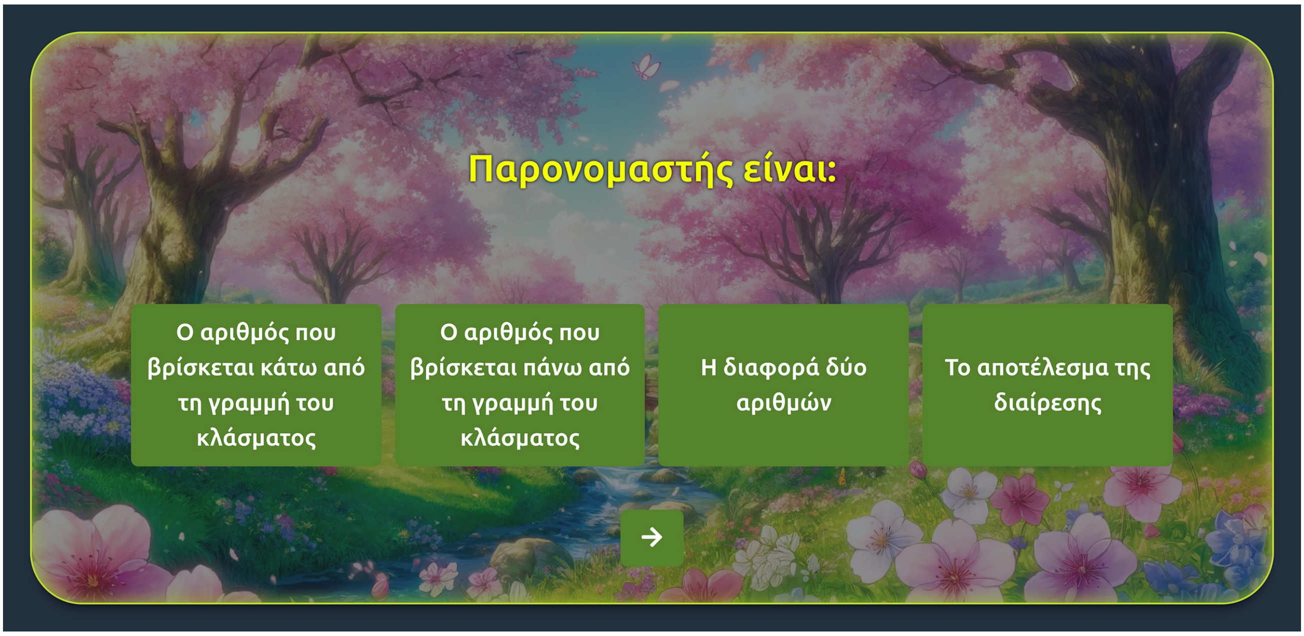

Fig. 2. A representative mathematical terminology comprehension task. The figure can be translated as: The denominator is:, with response options from left to right of (1) the number that appears below the fraction bar, (2) the number that appears above the fraction bar, (3) the difference between two numbers, and (4) the result of division.

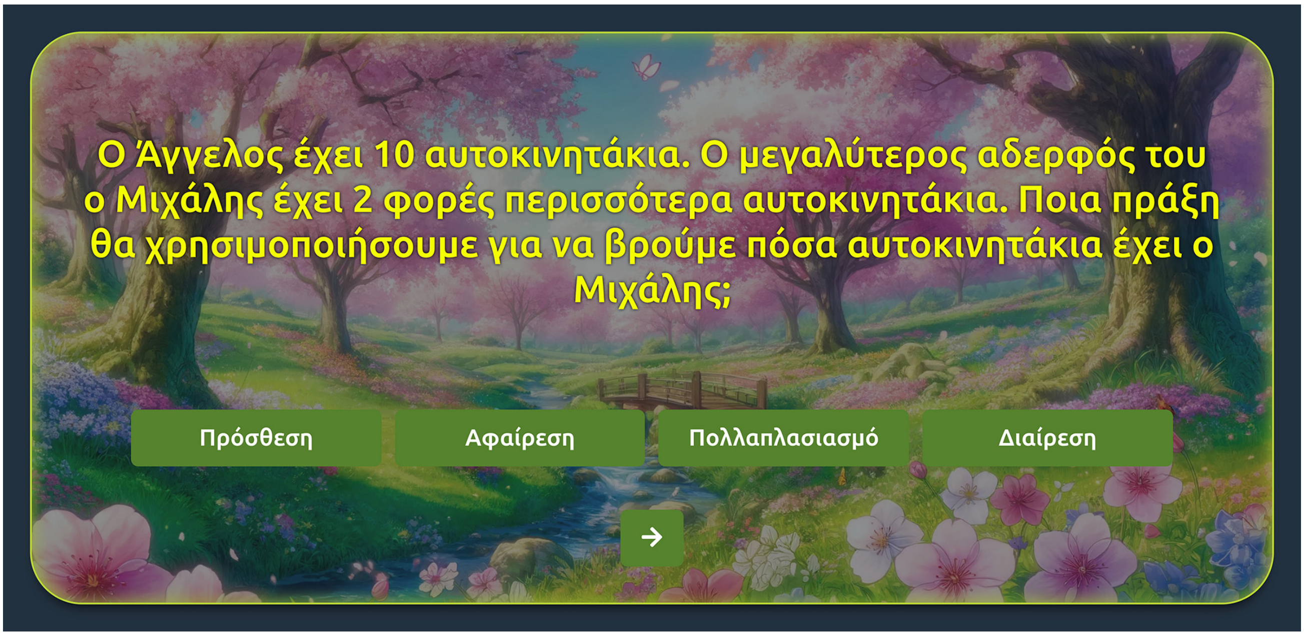

Fig. 3. A representative arithmetic word problem task. The figure can be translated as: A child has 10 cars. His older brother has twice as many cars as he does. Which operation should be used to determine how many cars the older brother has? Response options, from left to right, were (1) addition, (2) subtraction, (3) multiplication, and (4) division.

vidual screening indicators contributing to a broader multivariate neuropsychological profile, rather than as independent psychometric subscales.

Latency values were exported from the platform in minute-based units and were therefore interpreted as indicators of the time for task completion, rather than millisecond-level measures of participant reaction time.

Testing took place in an Informatics classroom under minimal adult supervision. Children completed the tasks on desktop computers using a mouse. The supervising teacher's only involvement was entering the application URL at the start of the session, after which the teacher sat at a distance from the child and did not provide feedback or performance support. Due to the potential for web-based latency to vary based on the device and browser being used, timing outcomes were interpreted conservatively and primarily at the group level. Sessions were discontinued and excluded from analysis if a child was unable to complete the initial training trial, which served as a brief check of task comprehension and basic mouse-use competency. The administration of the Askisi-MD screener required approximately 30 min per child. Parental/guardian consent was obtained for all participants.

2.3 Implementation

Participants completed exercises through the Askisi-MD neuropsychological web application, which recorded all results in a relational database. Once data collection was complete, the dataset was extracted for offline analysis, consisting of supervised machine learning models trained and evaluated to distinguish between children with and without MLD. Models used for classification included SVC, RF, and XGBoost.

2.3.1 Web Application Architecture

All task exercises were delivered through a Laravel PHP web application backed by a MariaDB (version 11.4 LTS; Redwood City, CA, USA) relational database. The application was designed for reliability and data integrity, ensuring that all participant-related data was securely stored.

Laravel's Model-View-Controller (MVC) architecture allowed for effective separation and facilitated the smooth, independent management of data models, server-side logic, and presentation. Application Programming Interface (API) endpoints provided controlled access to exercise data, enabling safe retrieval of results and session information.

The MariaDB database contained structured tables for participants, exercises, sessions, and results, with relationships that supported flexible grouping by test session. This structure allowed for the efficient aggregation of metrics, including task scores and summary statistics, providing a basis for subsequent machine learning analysis.

Security was a key consideration. Laravel's native features were leveraged to safeguard the application from security concerns such as Structured Query Language (SQL) injection and Cross-Site Request Forgery (CSRF). We enforced authenticated access for all API routes of the application in order to maintain the integrity of stored data and implemented measures to prevent Denial-of-Service (DoS) attacks, thus maintaining the application's availability and responsiveness even when facing heavy traffic through the use of Laravel's native rate-limiting and throttling capabilities.

All database interactions were handled on the server-side, ensuring that participant data remained confidential and tamper-proof while also validating the correct answers for students' posted results in order to avoid any potential for client-side request forgery while maintaining the integrity of the stored data.

2.3.2 Front-End Design

The web application used to complete the exercises was presented to participants through a user-friendly interface. Each task was presented clearly, with exercises loaded dynamically to ensure client-side responsiveness without overloading the browser with consecutive or large HTTP requests. All assets, includingCascading Style Sheets (CSS), JavaScript, images, and sounds, were hosted on a third-party content delivery network to leverage distributed caching and optimized delivery, thereby improving load times, responsiveness, and scalability.

The front-end design of the application was built using modern web standards, leveraging HTML5 and CSS3 for structure and styling. All exercises followed a consistent User Interface (UI) layout, and each type of element was positioned uniformly on the screen to improve usability and reduce cognitive load.

To further enhance the User Experience (UX), a small transitional screen was introduced, consisting of a snake-like path filled with bright colors to reflect progress between exercises. This gamification element was designed to engage and maintain the motivation of participating children.

2.3.3 Machine Learning Pipeline

Once data collection was complete, the recorded exercise results were extracted from the MariaDB database for offline analysis.

2.3.4 Dataset Description

The feature space presented to the classification models consisted of task-level scores for six tasks and latency indicators for four tasks, yielding 10 total features per participant. The classification models operated on a multivariate feature matrix rather than on individual task features. The target variable was a binary label distinguishing participants with MLD from control students without MLD. Students with MLD were assigned the class label of one (1), while students in the control group were assigned the class label of zero (0). A sample dataset for the six tasks is presented in Table 1.

2.3.5 Preprocessing

Before modeling, the dataset underwent preprocessing to ensure the input was clean and relevant. Non-descriptive columns, such as participant age and gender, were removed from the feature space, retaining only the exercise-related metrics for further analysis. This ensured that the model was only trained on the predictive signals relevant to MLD without including any potential demographics-related signal pollution. A binary target label was employed for study participants. Preprocessing was confined to this column removal step. No distributional preprocessing such as

Table 1. Sample dataset for the six tasks.

Tasks	Control group (Class 0)		MLD group (Class 1)	
	Task score	Time taken (min)	Correct answers	Time taken (min)
Go/No-Go error score	0	-	4	-
Visual pattern recognition abilities task	9	2.22	0	1.59
Visual working memory task	5	-	3	-
Mental calculation task	4	0.44	2	0.53
Mathematical terminology comprehension task	2	0.10	0	0.21
Arithmetic word problem task	2	0.22	1	1.01

MLD, Mathematical learning difficulties. Note: For the Go/No-Go task, the task score represents the number of errors, with higher scores indicating poorer inhibitory control. For all other tasks, higher task scores indicate better performance.

feature scaling, imputation, or outlier treatment was employed at any stage of the machine learning pipeline. Column removal does not constitute a source of data leakage as it carries no statistical information derived from the full dataset. Feature standardization was not applied in the present machine-learning pipeline. This is an important methodological limitation for the SVC model because SVC performance is sensitive to differences in feature scale. Therefore, the comparison between SVC and the tree-based models should not be interpreted as a fully optimized head-to-head model comparison. The absence of feature scaling may have placed SVC at a disadvantage relative to Random Forest and XGBoost, which are less sensitive to feature scaling. Accordingly, the SVC results should be interpreted as conservative and primarily exploratory.

2.3.6 Initial Train-Test Split

The dataset was randomly split into a training set and a test set while maintaining class balance through stratification, ensuring that both sets contained a proportional representation of participants with and without MLD. This split was employed to assess model performance on unseen data by fitting the model using data from the training set, while the test set served as an independent benchmark to report metrics such as accuracy, precision, recall, and F1 scores, thereby providing insight into the generalization ability of developed models.

2.3.7 Cross-Validation

To ensure correct evaluation of classification performance and reduce variance associated with a single train-test split, Stratified K-Fold cross-validation was employed. Stratification at this stage ensured that the proportion of participants with and without MLD remained consistent across folds, preserving class balance and avoiding skewed results. A 4-fold split was selected, providing multiple independent evaluations of model performance while maintaining sufficient sample size per fold.

Hyperparameter optimization was conducted via GridSearchCV within the training portion of each outer fold exclusively, using a 2-fold stratified inner cross-validation loop. This nested structure ensured that hyperparameter selection was performed without any exposure to the held-out evaluation fold, preventing information leakage between model selection and performance estimation. The best estimator identified by GridSearchCV was then evaluated on the held-out test set, resulting in performance metrics that had been computed on unseen partitions.

2.4 Data Analysis

Before comparing data between groups, the dataset was screened to identify missing values, impossible values, outliers, and instances of extreme task completion times. No missing values were identified across the Askisi-MD task scores or latency variables; therefore, no imputation procedure was required. Accuracy-based task scores were checked against the valid scoring range for each task, and no impossible values were detected. These accuracy variables were not Winsorized, as they represented fixed-range performance scores and extreme values could reflect genuine clinical variability. Latency variables were treated separately because minute-based task-completion time measures are more vulnerable to extreme observations, particularly in web-based assessments administered to child populations. To reduce the disproportionate influence of unusually short or long completion times, latency values were screened within MLD and control group using a group-specific mean ±3 SD criterion (see **Supplementary Table 2**). Values falling beyond these thresholds were Winsorized to the nearest acceptable boundary rather than removed from the dataset. This approach was selected to limit the influence of extreme latency values while preserving the full sample for analysis. It should be clarified that Winsorized latency values were used only for Welch's group comparison analyses. Winsorized latency values were not used in the internal coherence analyses, ROC analysis, logistic regression classification, cross-validation, SVC, Random Forest, XGBoost, or SHapley Additive exPlanations (SHAP) analysis. All internal coherence and predictive analyses were conducted using the raw task score and raw minute-based task-completion time variables exported from the Askisi-MD platform. This approach avoided using diagnostic group information during preprocessing and reduced the risk of label-informed data leakage. After this preprocessing step, the dataset was used for the main statistical analyses.

Feasibility indicators were summarized using descriptive statistics, and all analyses were conducted in SPSS 27.0. (IBM Corp. (2020). IBM SPSS Statistics for Windows (Version 27.0). Armonk, NY, USA) Group-level differences were tested through two-tailed independent-samples Welch's t-tests comparing the performance of participants with MLD to that of those in the control group.

Because group differences were examined across multiple cognitive and mathematical outcomes, the inflation of Type I error was controlled by subjecting p-values to Benjamini–Hochberg False Discovery Rate (FDR) correction with a target rate of $q = 0.05$ across the full set of comparisons. Additionally, Hedges' g was computed for each outcome to quantify the magnitude and direction of group differences using a metric that is comparable across measures, thus facilitating the more intuitive interpretation of these results beyond statistical significance alone.

Internal coherence was examined using McDonald's omega (Ω) for selected theoretically defined composites derived from the Askisi-MD task scores and latency indicators. The analyzed composites were the neuropsychological task-score composite, the total task-score composite, the mathematical processing-efficiency composite, the broader processing-efficiency composite, and the total score-and-latency composite. McDonald's Ω was selected because it provides a flexible estimate of internal coherence and does not require the restrictive assumption that all indicators contribute equally to the underlying construct.

Because several mathematical tasks consisted of only 2–4 items, reliability was not interpreted at the level of each mathematics subtest. Instead, McDonald's Ω was calculated for theoretically defined composites that combined conceptually related task indicators. Test–retest reliability was not available because the present study used a single-administration design. Consequently, subtest-level standard errors of measurement were not calculated.

Binary logistic regression analysis was also performed in SPSS 27.0 to examine the extent to which the Askisi-MD neuropsychological web-based screener can predict participating students' membership in the MLD or control groups. Classification performance was further analyzed using group-specific performance metrics that were calculated to evaluate the screener's accuracy for each group. Receiver operating characteristic (ROC) curve analyses were employed to assess the screener's ability to discriminate between children with MLD and controls, providing an overall measure of screening performance based on the area under the ROC curve (AUC).

Finally, post-hoc SHAP (SHapley Additive exPlanations) analysis was performed to calculate the contribution of individual cognitive and mathematical features to model predictions.

3. Results

3.1 Between-Group Comparisons of Askisi-MD Performance

Welch's independent-samples t-tests were initially used to compare the performance of children with MLD and typically achieving children on the Askisi-MD task scores and latency indicators. All analyses were subjected to Benjamini–Hochberg FDR correction for multiple comparisons, and Hedges' g was calculated to estimate effect sizes (Table 2).

Children with MLD also showed significantly longer minute-based task-completion times than controls for two of the four recorded completion-time indicators. Specifically, completion time was significantly longer for children with MLD on the Mental Calculation Task and the Arithmetic Word Problem Task after FDR correction ($p < 0.001$). In contrast, no significant group differences were observed for Visual Pattern Recognition Abilities Latency or Mathematical Terminology Comprehension Latency ($p > 0.05$). This pattern suggests that completion-time differences were more selective than task-score differences and were mainly observed in tasks requiring active mathematical processing, calculation, and arithmetic word-problem solving.

Latency was not recorded for the Go/No-Go Task or the Visual Working Memory Task due to the design of the web-based screening application. For the Visual Working Memory Task, latency was not included because task-completion time would have been influenced by each child's recall span and task progression. Specifically, not all children were exposed to the full set of number sequences, as the termination of the task was based on each participant's performance. As such, latency would not have been a reliable or comparable processing-efficiency index across participants. For the Go/No-Go Task, all children were given the same fixed response interval after the presentation of the complete set of stimuli. As such, the task was designed to capture accuracy-based inhibitory control performance rather than individual differences in task-completion time.

3.2 Internal Coherence of Selected Askisi-MD Composite Indicators

Next, McDonald's omega analysis was conducted to examine the internal consistency of theoretically defined Askisi-MD composites. Given the brief structure and restricted score range of several mathematical tasks, reliability was examined at the composite level rather than for each subtest separately. The mathematics tasks, which included two to four items, were therefore treated as screening indicators contributing to the overall Askisi-MD profile rather than as independent psychometric scales. The analysis used the raw task scores and raw task-completion times, measured in minutes, exported from the Askisi-MD platform. The Go/No-Go error score was reverse-coded before the analysis so that higher scores consistently indicated better

Table 2. Comparison of Askisi-MD task performance between children with MLD and control.

Askisi-MD tasks	Control group		MLD group		Welch's t	df	p	FDR-adjusted p	Hedges' g
	M	SD	M	SD					
Go/No-Go error score	0.02	0.15	0.84	2.08	6.39	265.74	<0.001	<0.001	0.555
Visual Pattern Recognition Abilities Task	7.37	2.10	6.20	2.43	−5.92	515.18	<0.001	<0.001	−0.514
Visual Pattern Recognition Abilities Latency	2.61	1.28	2.58	1.14	−0.28	519.10	0.776	0.776	−0.025
Visual Working Memory Task	3.93	1.74	2.69	1.48	−8.82	512.80	<0.001	<0.001	−0.767
Mental Calculation Task	2.99	0.98	2.35	1.08	−7.13	521.11	<0.001	<0.001	−0.620
Mental Calculation Latency	0.88	0.74	1.17	0.81	4.30	521.76	<0.001	<0.001	0.373
Mathematical Terminology Comprehension Task	0.92	0.71	0.42	0.58	−8.86	505.86	<0.001	<0.001	−0.770
Mathematical Terminology Comprehension Latency	0.31	0.28	0.33	0.24	0.88	513.98	0.379	0.421	0.077
Arithmetic Word Problem Task	1.63	0.57	0.88	0.81	−12.30	472.18	<0.001	<0.001	−1.069
Arithmetic Word Problem Latency	0.20	0.15	0.54	0.43	12.13	326.07	<0.001	<0.001	1.054

Note: Winsorized latency values were used exclusively for the present group-comparison analysis.
Welch's independent-samples t-tests indicated that children with MLD showed significantly poorer performance ($p < 0.001$) across all six task-score indicators. For five tasks, poorer performance was reflected by lower accuracy scores. For the Go/No-Go task, poorer performance was reflected by a higher error score, indicating more inhibitory control and response selection errors. MD, Mathematical Difficulties; FDR, False Discovery Rate.

Table 3. McDonald's omega analysis of task-score and completion-time composite indicators.

Composite	Indicator type	Variables included	McDonald's Ω
Neuropsychological task-score composite	Scores	Go/No-Go, Visual Pattern Recognition Abilities, Visual Working Memory	0.671
Total task-score composite	Scores	Go/No-Go, Visual Pattern Recognition Abilities, Visual Working Memory, Mental Calculation, Mathematical Terminology Comprehension, Arithmetic Word Problem	0.686
Mathematical processing-efficiency composite	Latencies	Mental Calculation Latency, Mathematical Terminology Comprehension Latency, Arithmetic Word Problem Latency	0.809
Broader processing-efficiency composite	Latencies	Visual Pattern Recognition Ability Latency, Mental Calculation Latency, Mathematical Terminology Comprehension Latency, Arithmetic Word Problem Latency	0.754
Total score-and-latency composite	Scores + latencies	All six task scores and four latency indicators	0.645

performance across all task-score variables. McDonald's total omega was calculated using standardized indicators for the full sample of 564 children in SPSS version 27.

The results are presented in Table 3.

As shown in Table 3, the neuropsychological task-score composite exhibited modest internal coherence (Ω = 0.671). The total task-score composite, which included all six task scores, also showed modest internal coherence (Ω = 0.686). The strongest internal coherence estimate was observed for the mathematical processing-efficiency composite, which included Mental Calculation Latency, Mathematical Terminology Comprehension Latency, and Arithmetic Word Problem Latency (Ω = 0.809), indicating good internal coherence. The broader processing-efficiency composite, which also included Visual Pattern Recognition Abilities Latency, showed modest to good internal coherence (Ω = 0.754). Finally, the total score-and-latency composite showed moderate to modest internal coherence (Ω = 0.645).

The Ω coefficients reported in Table 3 should be interpreted as evidence for the internal coherence of selected Askisi-MD composites, not as evidence that each brief mathematics task is a reliable standalone psychometric subscale. The strongest internal coherence was observed for the mathematical processing-efficiency composite, suggesting that latency indicators across mathematics-related tasks captured a relatively consistent dimension of task-completion efficiency. However, the individual mathematics tasks themselves were intentionally brief and should be viewed as screening probes within a broader multivariate profile.

3.3 Discriminant Validity and ROC Analysis

To test the discriminant validity of the Askisi-MD neuropsychological web-based screener, binary logistic regression analysis was next performed, calculating sensitivity and specificity indicators for the screener to examine the

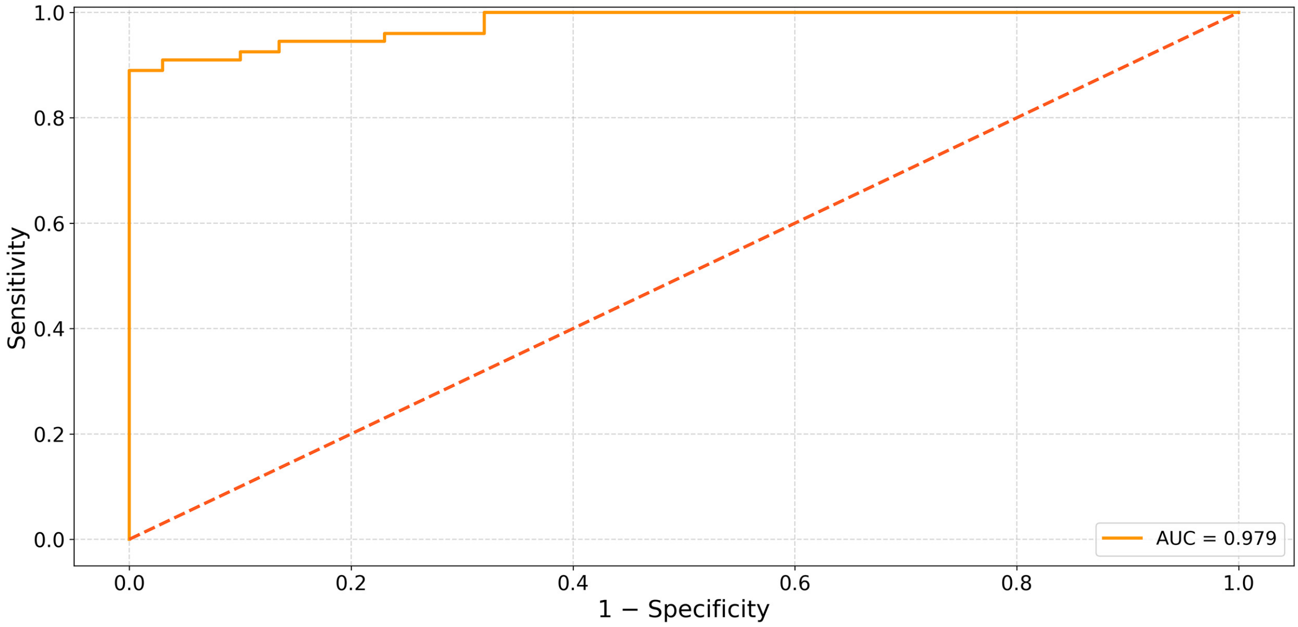

Fig. 4. Receiver operating characteristic (ROC) curve for Askisi-MD neuropsychological web-based screener.

percentages of students correctly assigned to the MLD and control groups. The overall ability of this screener to distinguish between these two groups of participants was assessed based on the AUC for the ROC curve. Good discriminative performance was observed, with an AUC of 0.979 (95% CI: 0.967–0.991), indicating that the screener reliably distinguished between children with MLD and controls (Fig. 4).

Classification performance was additionally evaluated using a classification matrix summarizing the correct and incorrect classifications for each group. From this matrix, group-specific indices were computed to separately assess the screener's accuracy for children with MLD and for controls (Table 4).

Table 4. Classification metrics.

Metric	Control group	MLD group	Overall
Accuracy	—	—	0.93
Precision	0.92	0.94	—
Recall	0.95	0.91	—
F1 score	0.93	0.92	—

The classification metrics presented in Table 4 indicate that the Askisi-MD neuropsychological screener distinguished children with MLD from typically developing controls with strong and balanced performance. The analysis included all 564 children: 282 children with MLD and 282 typically developing controls. MLD was defined as the positive class, and classifications were based on a predicted-probability threshold of 0.50. The model correctly classified 257 children with MLD and 267 controls. In contrast, 15 controls were incorrectly classified as having MLD, while 25 children with MLD were incorrectly classified as controls. The overall classification accuracy was 0.93. For the MLD group, precision was 0.94, recall (sensitivity) was 0.91, and the F1 score was 0.92. For the control group, precision was 0.92, recall (specificity) was 0.95, and the F1 score was 0.93. These findings indicate that the model classified both groups accurately without favouring one group over the other. The group-specific indices are consistent with strong performance for both classes. For the MLD group, the precision was 0.94, indicating that when the analysis identified a child as having MLD, this classification was correct in most cases. Recall for children with MLD was 0.91, meaning that the screener successfully identified about 91% of children with MLD. The corresponding F1 score of 0.92 reflects a strong balance between precision and recall, supporting the screener's reliability when prioritizing correctly detecting MLD while minimizing instances of misclassification.

In the control group, the precision was 0.92, while the recall was 0.95, indicating that the model was also able to reliably recognize typically developing children, rarely mislabeling them as being affected by MLD. The F1 score of 0.93 further suggests consistent performance for the control class. Taken together, these results emphasize that the high accuracy of the classifier is not due to it favoring one group at the expense of the other; rather, it provides a high degree of accuracy for both MLD and control subjects, supporting the discriminant validity of the Askisi-MD neuropsychological screener.

3.4 Machine-Learning Classification Performance

Furthermore, classification analyses were performed using three supervised machine-learning models: SVC, RF,

Table 5. Cross-validation classification performance of machine learning models used to discriminate between MLD and control groups.

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Mean average
SVC	Accuracy	0.87	0.88	0.84	0.87	0.87
	Precision	0.88	0.88	0.88	0.90	0.89
	Recall	0.88	0.86	0.77	0.82	0.83
	F1 score	0.86	0.87	0.82	0.86	0.85
RF	Accuracy	0.91	0.84	0.91	0.89	0.89
	Precision	0.87	0.84	0.92	0.90	0.88
	Recall	0.94	0.82	0.89	0.86	0.88
	F1 score	0.91	0.83	0.91	0.88	0.88
XGBoost	Accuracy	0.87	0.91	0.92	0.92	0.91
	Precision	0.91	0.87	0.94	0.94	0.92
	Recall	0.80	0.94	0.89	0.89	0.88
	F1 score	0.85	0.91	0.91	0.91	0.90

SVC, Support Vector Classifier; RF, Random Forest; XGBoost, eXtreme Gradient Boosting.

and XGBoost. All models achieved strong predictive performance. XGBoost presented the highest mean accuracy, precision, and F1 score, while XGBoost and Random Forest achieved the same mean recall. The cross-validated classification metrics for each model across the four-fold stratified cross-validation process are presented in Table 5. Metrics include accuracy, precision, recall, and F1 score for each fold, as well as the mean average value across all folds. The XGBoost model achieved the highest mean performance across metrics, indicating superior discriminative ability relative to the RF and SVC models.

Model discrimination was further examined using ROC analysis. Fig. 5 presents the XGBoost ROC curve, which was constructed by combining out-of-fold predictions across all cross-validation folds. XGBoost showed the strongest mean cross-validated performance, achieving an AUC of 0.956 and demonstrating strong discrimination between children with MLD and typically developing controls.

The AUC derived from the logistic regression analyses in Fig. 4 reflects performance under the logistic model fitted to the present dataset and should be interpreted as model-based discrimination in this sample. In contrast, the machine learning results in Fig. 5 were evaluated under stratified cross-validation, which provides a more conservative estimate of generalization because predictions are generated on held-out folds. Accordingly, the two AUC values are not directly identical, with logistic regression analysis supporting parametric, interpretable inferential validation, whereas cross-validated machine learning enables the evaluation of predictive stability under resampling.

3.5 Explainable Machine-Learning Analysis

Post-hoc SHAP analysis was conducted on the final XGBoost model to interpret the contributions of individual features to model predictions (Fig. 6). SHAP values were computed on out-of-fold predictions from the cross-validation procedure to reduce optimism in attribution. Feature rankings were inspected across folds to evaluate the stability of the importance pattern. Because several screener measures are conceptually correlated with one another (e.g., mathematical language and the selection of word problem operations), SHAP attributions were interpreted as model-conditional contributions that may distribute credit across correlated predictors rather than isolating a single causal construct.

The SHAP summary plot (Fig. 6) indicated that both mathematical tasks (Arithmetic Word Problem, Mathematical Terminology Comprehension, Mental Calculation) and the cognitive task (Go/No-Go) demonstrated the largest absolute SHAP values, suggesting that both domain-specific mathematical performance and domain-general cognitive control contributed meaningfully to the model's classification decisions. More specifically, Arithmetic Word-Problem Task Latency showed the highest importance, followed by Arithmetic Word-Problem performance, Mental Calculation, Mathematical Terminology Comprehension, and Go/No-Go errors.

4. Discussion

The aim of this study was to develop and validate Askisi-MD as a web-based neuropsychological screening application capable of identifying school-age children with MLD. The Askisi-MD application revealed group differences between children with MLD and typically developing controls. Specifically, the MLD group consistently exhibited significantly worse accuracy across all six tasks ($p <$ 0.001). In analyses of minute-based task-completion time, children with MLD required significantly longer completion times on the Mental Calculation and Arithmetic Word Problem tasks ($p < 0.001$). Completion times did not dif-

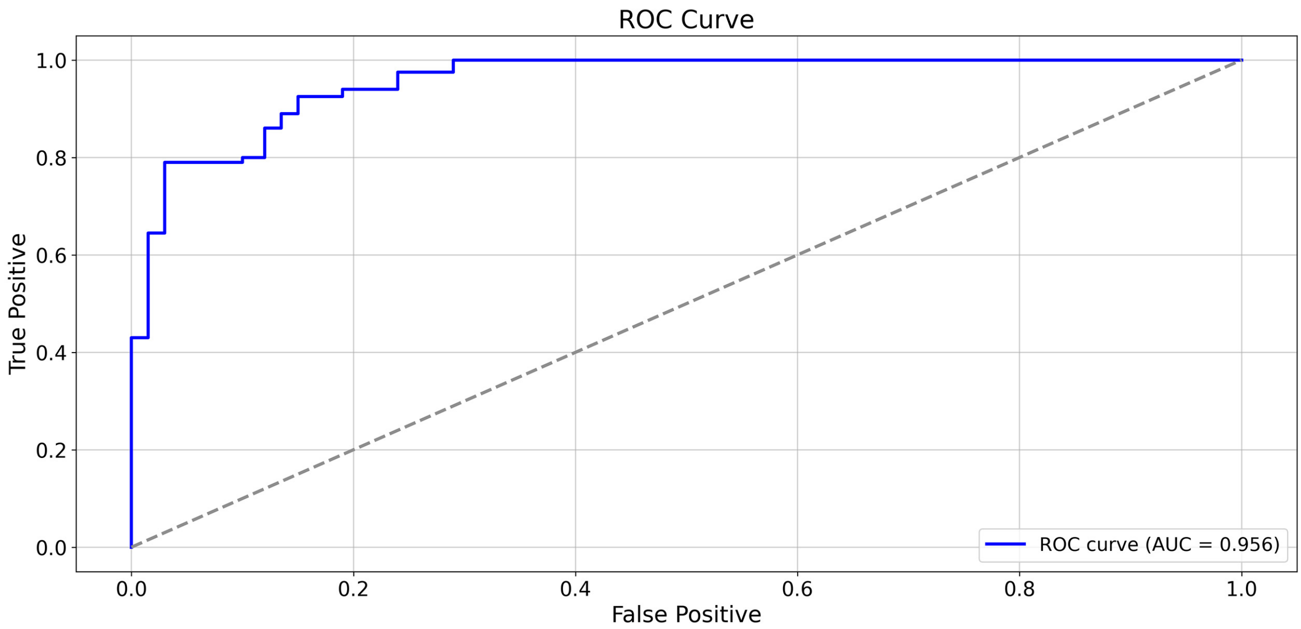

Fig. 5. ROC curve for XGBoost model (area under the ROC curve (AUC) = 0.956).

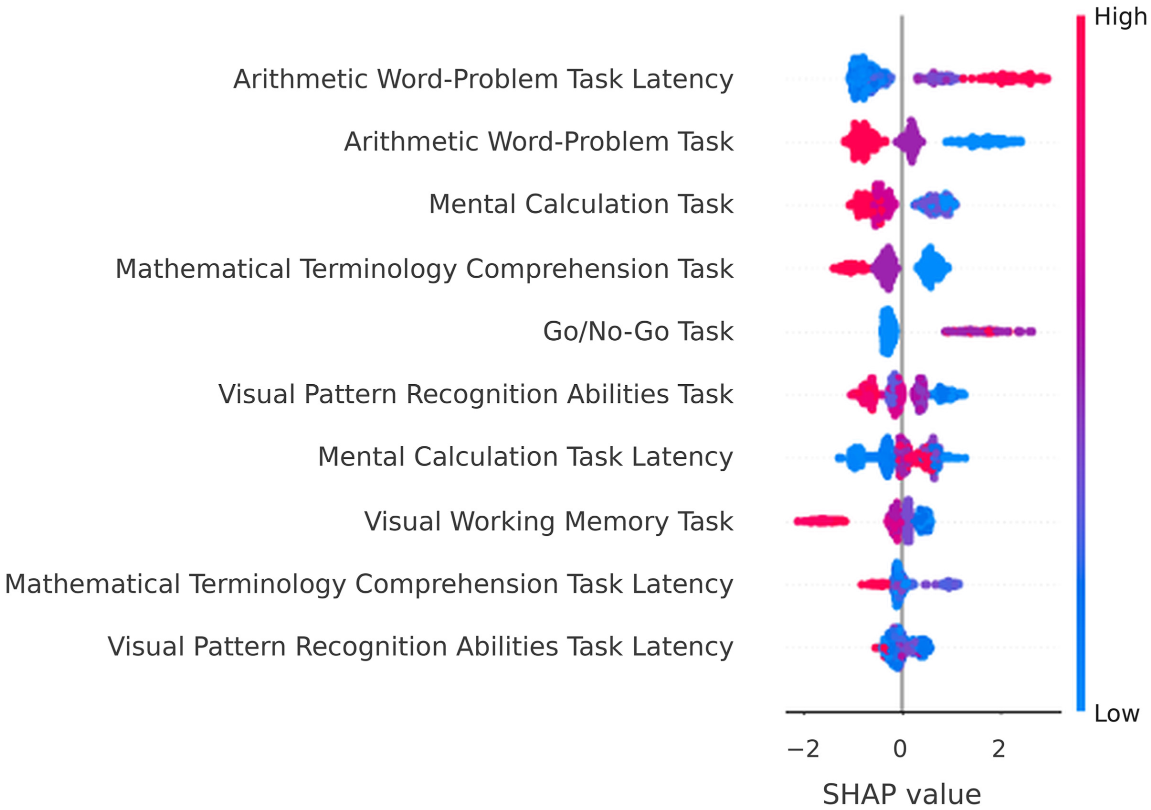

Fig. 6. SHAP summary plot for the XGBoost model, illustrating the magnitude and directionality of feature contributions to model predictions.

fer significantly between groups for Visual Pattern Recognition or Mathematical Terminology Comprehension ($p > 0.05$). The MLD group showed a slightly lower mean completion time for Visual Pattern Recognition and a slightly higher mean completion time for Mathematical Terminology Comprehension. Response latency was not recorded for the Go/No-Go and Visual Working Memory tasks. In addition, to minimize fatigue or discomfort and to ensure standardized administration across participants, the Visual Working Memory task was discontinued when a child produced two consecutive errors or three errors in total. Taken together, these findings support the first hypothesis of the study, which proposed that Greek children with an established MLD diagnosis derived from written testing at a state diagnostic center would show lower accuracy scores and longer minute-based task-completion time on Askisi-MD tasks than typically achieving controls. This pattern was confirmed across all six task score indicators, with poorer performance reflected by lower scores on five accuracy based tasks and by higher error scores on the Go/No-Go task and for two of the four latency-based indicators.

It should be noted that Winsorization was applied only for the group-comparison analysis. This procedure was used only for the latency variables. Its purpose was to reduce the influence of extreme minute-based task-completion times while preserving the full sample. Accuracy-based task scores were treated differently. They were only checked against their valid scoring ranges and were not Winsorized. This was because these variables had fixed minimum and maximum values, and extreme scores could reflect genuine clinical variability. Latency variables were more vulnerable to extreme values. In web-based assessment, completion times may be affected by hesitation, attentional fluctuation, temporary distraction, or technical delay. For this reason, latency values were screened within each group using a mean ±3 SD criterion. Values outside this range were Winsorized to the nearest acceptable boundary. This decision should be considered when interpreting the findings. Winsorization reduced the effect of unusually extreme latency values without excluding participants. Therefore, the group-level latency results provide a more stable estimate of differences in task-completion efficiency. However, this procedure does not remove all timing limitations associated with web-based assessment.

These findings are consistent with prior research using computerized measures to present mathematical and related cognitive processes in children with MLD or developmental dyscalculia. Studies employing computer-based screening batteries that combine timed number processing tasks have consistently reported performance differences between dyscalculia/MLD groups and typically developing peers, supporting the sensitivity of brief digital tasks as tools for case–control discrimination [63]. Similarly, large-scale online dyscalculia screeners designed for school-age samples have shown that children with weaker mathematical achievement perform significantly worse on core numerical and arithmetic components, demonstrating the feasibility of web-delivered approaches for differentiating groups based on ability [46,47]. In addition, computerized approximate number system (ANS) paradigms have been used to differentiate children with MLD/dyscalculia from controls in multiple cohorts, with reported group differences in ANS acuity or related indices [64,65,66,67]. Collectively, these reports suggest that computerized screening and assessment tools can detect group-level impairments in numerical processing and calculation, particularly on timed arithmetic and magnitude-processing tasks, paralleling the pattern observed for the Askisi-MD neuropsychological web application screener.

After FDR correction for multiple comparisons, the MLD group showed poorer performance across all six task-score indicators ($p_{FDR} < 0.001$). The magnitude of these differences was generally from negligible to large. ($g = 0.025$ to -1.069) for Visual Pattern Recognition, Visual Working Memory, Mental Calculation, and Mathematical Terminology Comprehension. For the Go/No-Go task, the positive effect size reflected the error-score direction, as higher scores indicated poorer inhibitory-control performance. The largest effect was observed for the Arithmetic Word Problem task ($g = -1.069$). This pattern suggests that the strongest group difference was found in higher-level mathematical reasoning, while sizeable differences were also present in executive control and visuospatial/working-memory processes that support mathematical performance.

Completion-time findings were more task-specific. Minute-based task-completion time did not differ significantly between groups for Visual Pattern Recognition Abilities Latency or Mathematical Terminology Comprehension Latency. This indicates that these tasks differentiated groups mainly through task-score performance rather than through completion-time indicators. In contrast, children with MLD showed significantly longer completion times on the Mental Calculation and Arithmetic Word Problem tasks. Taken together, these findings point to a broad task-score disadvantage across the Askisi-MD indicators, with longer completion times mainly when tasks required multi-step computation and integration of intermediate results. These results support the first hypothesis and show that Askisi-MD differentiated the MLD and control groups across both task-score and selected completion-time indicators.

These results are in the same line with prior work using computerized or tightly timed measures and group-comparison statistics to characterize MLD. Specifically, studies of executive function and mathematical difficulties have previously employed Go/No-Go tasks and reported group differences when using conventional group comparisons and effect-size indices, linking weaker inhibitory control to poorer mathematical outcomes [68]. The sizable group differences observed in visual working memory and complex mathematical performance in the present study align with evidence that working-memory limitations are positively correlated with MLD [69] and that these deficits are particularly apparent when tasks require maintaining and manipulating intermediate results [70]. MLD is often conceptualized as a multidimensional condition in which impaired mathematical performance arises from deficits in core number processing, memory, reasoning, or visuospatial functioning arising either alone or in combination [34].

This view emphasizes heterogeneity in neurocognitive profiles and suggests that children with MLD should not be regarded as constituting a single homogeneous group. The present results from the Askisi-MD application thus fit within the broader literature in this space, suggesting that the most pronounced differences in performance between children with MLD and typically developing controls manifest when completing tasks that combine numerical processing with executive/working memory demands, while differences in task-completion time are less consistent and may be most evident when multi-step computation or reasoning is required.

The McDonald's Ω results should be interpreted as evidence for the internal coherence of selected Askisi-MD composite indicators, rather than as evidence that each brief mathematics task is a reliable standalone psychometric subscale [71]. The neuropsychological task-score composite and the total task-score composite showed moderate internal coherence, suggesting that these grouped indicators can operate together as part of a broader multidomain screening profile. The strongest internal coherence was observed for the mathematical processing-efficiency composite, indicating that latency measures for mental calculation, mathematical terminology comprehension, and arithmetic word-problem solving may capture a relatively coherent dimension of mathematical task-completion efficiency. However, the individual mathematics tasks were intentionally brief, containing only 2–4 items, and should therefore be interpreted as short screening probes rather than independent achievement or diagnostic subscales.

This interpretation is consistent with previous evidence showing that mathematical difficulties are associated not only with lower accuracy, but also with differences in timed task performance, working memory, and other domain-general cognitive processes [38,72]. Kroesbergen et al. (2023) [40] similarly emphasized an association between MLD and weaknesses across both mathematics-specific and broader cognitive domains, supporting the need to assess multiple skill areas rather than relying on a single performance indicator [40]. Therefore, the stronger Ω coefficient observed for the mathematical latency composite may indicate that the time needed to complete calculation, numerical concept, and problem-solving tasks reflects a shared processing efficiency dimension within the MLD group.

The modest reliability of the broader processing-efficiency composite, which additionally incorporated visual-discrimination latency, further suggests that processing efficiency may not be restricted to mathematical operations alone. This visual-discrimination latency may reflect a more general perceptual-cognitive efficiency component that contributes to performance during mathematically relevant tasks. This interpretation is aligned with evidence that children with mathematical difficulties may show weaknesses in broader domain-general processes, including attention, executive function, working memory, and timed task performance [73]. However, the slightly lower Ω value for the broader latency composite indicates that visual-discrimination latency is related to, but not fully equivalent to, mathematical processing latency. This distinction is theoretically meaningful because mathematical task latency may involve number-specific processing, arithmetic retrieval, problem representation, and working memory demands, whereas visual-discrimination latency may depend more strongly on perceptual scanning, visual attention, and discrimination speed.

The score-based composites exhibited modest internal consistency, particularly when all neuropsychological and mathematical task scores were combined into a single total score. This is consistent with the multidimensional design of the Askisi-MD screener and with the broader literature suggesting that the impairments that characterize MLD are not uniform or homogeneous across cases [74]. Accuracy-based scores pertaining to inhibitory control, visual discrimination, working memory, calculation, numerical concepts, and problem solving do not represent a single narrow construct. Instead, they describe different components of the neuropsychological and mathematical profile for a given child. For example, calculation may depend more strongly on arithmetic procedures and fact retrieval, while numerical concepts may reflect number knowledge and quantitative understanding, and problem solving requires input from working memory, executive control, and the active manipulation of task-relevant information [75,76].

The moderate internal consistency observed for the total score and latency composite also supports this interpretation. Accuracy and latency provide complementary but non-identical information. Accuracy reflects whether the child reached the correct response, whereas latency reflects the efficiency and cognitive cost associated with reaching that response. This distinction is particularly relevant in children with MLD, as previous studies have shown that these children may exhibit higher error rates and longer task-completion times when completing cognitive and mathematical tasks [77]. As such, some children may respond accurately but slowly, whereas others may respond quickly but with reduced accuracy. Combining accuracy and latency into one broad composite may therefore obscure clinically meaningful differences in performance style. These findings support the interpretation of Askisi-MD as a multidomain neuropsychological screener whose indicators show composite-level coherence, rather than as a set of individually reliable mathematics subscales or as a focused mathematical achievement scale.

The latency-based mathematical processing-efficiency composite appears to be the most internally coherent indicator, suggesting that processing efficiency is a meaningful shared dimension in the MLD group. At the same time, the broader task scores remain clinically important, as they describe distinct neuropsychological and mathematical functions. This interpretation is consistent with contemporary evidence that mathematical learning difficulties involve both domain-specific mathematical weaknesses and domain-general cognitive processes, including working memory, timed task performance, attention, and executive function [38,73,75].

After conducting the reliability analysis, the next stage entailed a binary logistic regression analysis focused on the classification performance, sensitivity, and specificity of the Askisi-MD neuropsychological web-based screener. The aim of this analysis was not to treat the Askisi-MD indicators as a single unidimensional scale, but to examine whether the combined pattern of neuropsychological and mathematical task indicators comprising this model could be used to estimate the probability of group membership and to distinguish children with MLD from typically developing peers. The inclusion of raw task score and latency variables in the predictive analyses was important also for practical screening reasons. In practice, it maintained the conditions for the use of Askisi-MD by teachers and multidisciplinary school teams in everyday screening contexts. In real-world school-based screening, the diagnostic status of a child is unknown before assessment; therefore, the predictive models were built on raw observed performance indicators rather than group-adjusted latency values. The same principle was applied to the coherence analyses, which were also based on raw latency values rather than Winsorized latency values. This ensured that all model-based and composite-level analyses used the same observed task indicators that would be available in routine school-based screening. The results supported its ability to accurately distinguish between these groups, as the ROC curve indicated excellent discrimination. This finding suggests that, across a wide range of possible decision thresholds, the predicted probabilities generated by the binary logistic regression model were consistently higher for children with MLD than for typically developing children. In practical terms, the web-based screener exhibited a strong capacity to separate the two groups rather than relying only on a single, potentially arbitrary cut-off score.

Classification performance was further examined using a classification matrix and associated diagnostic indices. The overall accuracy indicated that the binary logistic regression model correctly classified most children. Importantly, performance was balanced across groups and did not reflect high overall accuracy driven by better classification in only one group. For children with MLD, precision indicated that when the model classified a child as having MLD, the classification was correct in most cases. Recall further indicated that the model detected approximately 9 out of every 10 children with MLD, while the F1 score reflected a strong balance between precision and recall. For the control group, the model also performed well, suggesting that typically developing children were rarely misclassified as having MLD. Taken together, this pattern of accuracy, precision, recall, and F1 scores indicates that the model's performance was not achieved by favoring one group at the expense of the other. Instead, the findings support the discriminant validity of Askisi-MD as a multidomain web-based screening tool for identifying children at risk for MLD. These findings provide support for the second hypothesis of the study, which proposed that task-derived features from Askisi-MD would assist in distinguishing children with MLD from controls. The ROC analysis indicated high discriminative performance and the threshold-based classification indices suggested above-chance assignment of children to the MLD and control groups.

This approach to evaluating the Askisi-MD screener is consistent with analytic strategies used in previous research on educational and mathematical-risk screening. Studies developing or validating screening instruments commonly use ROC curve analysis to evaluate discrimination, examine possible cut-off scores, and report classification indices such as sensitivity, specificity, accuracy, precision, recall, and F1 score to determine how effectively the tool identifies children who may require further assessment or support [78,79,80]. In this context, the high AUC observed for Askisi-MD, together with the balanced precision and recall across groups, suggests that the screener can perform as a binary classification tool. These findings support the second hypothesis of the present study, indicating that the Askisi-MD neuropsychological web-based screener can successfully discriminate between children with MLD and their typically developing peers.

As noted above, in the present study, supervised machine learning classification was used to evaluate the classification performance of the web-based neuropsychological screening battery as a means of distinguishing children with MLD from typically developing peers. The task indicators were treated as multidomain neuropsychological and mathematical predictors rather than as interchangeable components of a single total score. This approach was appropriate, as the Askisi-MD battery was designed to capture different but clinically relevant aspects of performance, including inhibitory control, visual pattern recognition, visual working memory, calculation, mathematical terminology comprehension, arithmetic word problem solving, and processing efficiency.

Across a four-fold stratified cross-validation procedure, all three models produced effective and relatively stable performance, with mean accuracy ranging from 0.87 for the SVC model to 0.91 for the XGBoost model. The pattern of results suggests that the screener contains features with meaningful discriminative information rather than relying on distinctive fold-specific variance. This is particularly important in screening research, where a strong emphasis is placed on generalizability over maximal fit to a single sample [68]. In this context, the present classification findings suggest that the heterogeneous task indicators that comprise Askisi-MD can offer clinically meaningful information suitable for distinguishing children with MLD from typically developing peers [81].

Although the SVC and RF classifier models achieved high precision and recall overall, the XGBoost model exhibited the most consistent advantage across cross-validated metrics. This finding is plausible given that gradient-boosted decision tree ensembles are designed to iteratively correct residual errors and can capture nonlinear interactions among predictors while controlling for overfitting through regularization and shrinkage [82]. In contrast, linear or kernel-based margin classifiers can be highly effective in many diagnostic contexts [82], but their performance may be constrained when class separation depends on complex combinations of cognitive and academic features rather than on a single dominant dimension. Similarly, RF models benefit from aggregation across many decorrelated trees [83], while boosting can sometimes offer incremental gains when subtle, feature-dependent decision boundaries are present [82]. Importantly, the differences among models in the current results appear modest in absolute terms but consistent in directionality, which supports reporting multiple classifiers rather than relying on a single algorithm. These results support the third hypothesis of this study, as the utilized machine learning models provided classification performance comparable to that of simpler linear approaches. In addition to fold-wise classification metrics, ROC analysis offered an additional model-agnostic index of discrimination, with the ROC curve for the XGBoost model exhibiting strong differentiation between groups. These findings support the third hypothesis of the study, suggesting that nonlinear machine-learning models captured meaningful multivariate patterns within the Askisi-MD task-derived features and provided good classification performance in distinguishing children with MLD from typically achieving controls.

This approach aligns recommendations to complement point estimates of accuracy with threshold-sensitive indices that characterize the tradeoff between sensitivity and specificity, particularly for screening instruments where the costs of false negatives and false positives may differ across settings [84]. When performing digital testing for MLD and related learning disorders, ROC-based evaluation has been increasingly used as a validation component. The BM-PROMA multimedia battery for mathematics-related cognitive and basic skills reported ROC analyses with promising discrimination indices for the identification of children at risk for dyscalculia [85]. In addition, a gamified computerized screening and intervention framework covering dyslexia, dysgraphia, and dyscalculia previously incorporated ROC curves for dyscalculia prediction, reflecting a similar emphasis on classification-oriented validation rather than group comparisons alone [86]. More broadly, recent reviews have explored growing interest in AI-supported dyscalculia screening and the use of standard machine learning classifiers in educational and clinical screening pipelines, while also emphasizing heterogeneity in task design and validation rigor across tools [87].

One notable strength of the present study is that model interpretability was explicitly examined through SHAP analysis of the final XGBoost model. Explainability is increasingly regarded as important in the context of neuropsychological and educational screening because it helps connect model decisions to theoretically meaningful constructs while reducing the risk that classifiers rely on spurious correlations [88].

SHAP analysis was used after model training to make the XGBoost results easier to interpret [88]. The aim was to examine which Askisi-MD indicators had the greatest influence on the model's predictions. This is useful because complex machine-learning models can classify children accurately, but their decision process is not always easy to understand in applied screening settings. In this study, the strongest SHAP values were observed for arithmetic word problem latency and performance, mathematical terminology comprehension, mental calculation, and inhibitory-control errors. This suggests that these indicators had greater influence on the XGBoost predictions than the other task variables.

However, SHAP values do not prove that these tasks represent separate neuropsychological mechanisms. They also do not demonstrate causal effects. They only describe how each variable contributed to the predictions of this specific trained model [88]. Some variables with higher SHAP values also showed stronger group differences in the conventional analyses. This provides a useful descriptive comparison. However, it should not be treated as independent proof that the model has identified the psychological meaning of MLD. It only suggests that the model partly relied on task indicators that also differentiated the MLD and control groups in this sample. This pattern is consistent with evidence that MLD often involves both domain-specific weaknesses in mathematical and numerical processing and domain-general contributors, such as working memory, timed task performance, attention, and executive control [38,74]. Word problem performance has also been linked to working memory and language-related demands, supporting the cautious interpretation that arithmetic word problem tasks may reflect both mathematical reasoning and broader cognitive resources [89]. Several Askisi-MD variables are conceptually and statistically related, so SHAP importance may be shared across correlated predictors. Future studies should test the model in independent samples before stronger conclusions are made about the stability, fairness, or clinical meaning of the feature-importance pattern.

Additionally, these machine learning and ROC findings extend the validation of the screening battery beyond mean group differences by demonstrating discriminative performance under cross-validation and describing which task components had greater influence on model predictions.

This combination of predictive validation, including cross-validated classification metrics and AUC values, together with SHAP-informed model interpretability, is consistent with current recommendations for reporting and evaluating predictive models, as discrimination, calibration, transparent reporting, and interpretability are essential for judging practical utility in real-world settings [90]. Future work should examine model calibration by clarifying whether predicted probabilities align with observed risk, exploring optimal decision thresholds for use in schools versus clinical referral contexts, and conducting external validation in new participant cohorts. These steps are crucial to inform the practical implementation of screening tools that initially show promise in a selected cohort [90,91,92].

In neuropsychological terms, Askisi-MD is most appropriately conceptualized as a first-line screening tool, as it employs a structured and time-feasible approach to sampling the mathematical decision-making of those completing the screener, providing an opportunity to identify children with potential MLD who may require comprehensive assessment. When the screener yields a positive result consistent with potential MLD, teacher interviews should be conducted, and curriculum-based work samples should be reviewed, followed by standardized mathematical achievement testing and evaluation of comorbid attention/reading difficulties, when indicated. Thresholds for MLD positivity can also be adjusted contextually, as school-wide screening may prioritize sensitivity to reduce missed cases, whereas clinical intake may favor specificity to reduce unnecessary downstream testing.

Limitations

Several limitations of this study warrant consideration. For one, the study employed a clinically-defined MLD that was compared to age- and gender-matched typically developing peers. This case–control approach is useful for establishing initial discriminant validity, but may yield stronger separation than would be expected in community settings, as real-world mathematical difficulties vary in severity and comorbid conditions are common.

Furthermore, latency was not recorded for the Go/No-Go and Visual Working Memory tasks because of the way these tasks were implemented. This limited the evaluation of completion-time–accuracy relationships across the full screener. In addition, the Visual Working Memory task included an early-termination rule to reduce fatigue and standardize administration. This rule may have reduced variability and comparability across participants, particularly among children with lower performance.

The diagnostic documentation used to identify children with MLD may also warrant further methodological consideration. All children in the MLD group had been previously diagnosed by official Greek public diagnostic centers, and the research team relied on these existing multidisciplinary records for group classification. The detailed reports prepared by the multidisciplinary team included evidence of functional academic difficulties in mathematics, school-based impairment, developmental and educational information, behavioral observations, and exclusionary considerations consistent with DSM-5 descriptions of Specific Learning Disorders [14]. As such, MLD group classification was not based solely on an IQ–achievement discrepancy criterion. The utilized diagnostic records also included WISC-III-based information and discrepancy-oriented interpretations, reflecting the assessment procedures available in Greek clinical practice at the time of diagnosis. The research team did not administer the WISC-III [60] and did not independently apply the discrepancy criterion. Nevertheless, because it was not possible to reassess the participating children with a fully updated diagnostic protocol, this reliance on pre-existing diagnostic documentation should be acknowledged. Future studies may achieve greater diagnostic precision by documenting DSM-5 criteria [14] and by using updated cognitive and academic assessment tools where available.

Another limitation is the absence of a clinical control group. The present study compared children with MLD to typically developing controls, which is appropriate for initial case–control validation but does not establish whether Askisi-MD can distinguish MLD from other neurodevelopmental or learning disorders, such as dyslexia, ADHD, or mixed learning difficulties. Therefore, the present findings support initial discriminant validity between MLD and typical development, but not disorder-specific differential classification. Future studies should include clinical comparison groups to examine whether the screener can differentiate MLD from other learning and attentional profiles. The exclusion of children who were not native Greek speakers or who lived in bilingual or multilingual family environments also limits generalizability. This criterion was used to reduce linguistic variability during initial validation of a Greek-language screening tool, particularly for tasks involving mathematical terminology and word problem comprehension. However, the findings should not be generalized to bilingual or multilingual children without further validation.

Future research should examine measurement performance across linguistically diverse groups and evaluate whether task adaptations are needed for bilingual learners. Additionally, although model performance was evaluated using repeated splits and cross-validation, all analyses were conducted within the same overall dataset. Independent replication in a new cohort and testing under typical school conditions are needed before firm conclusions can be drawn about generalizability in routine screening practice. Moreover, the brevity of some subtests may limit individual-score stability and internal consistency estimates. Future iterations will expand item pools and evaluate reliability while retaining the short administration window. Finally, the present work focused on discrimination between groups, but additional validation steps would strengthen practical implementation. Future studies may benefit from examining risk score calibration, employing decision thresholds tailored to different referral contexts, and examining performance across sociodemographic subgroups to ensure that screening decisions are equitable.

5. Conclusions

Askisi-MD is a web-based neuropsychological screener that integrates cognitive measures with curriculum-relevant mathematical tasks. Across five tasks in this screener, children with MLD exhibited significantly lower performance than typically developing controls ($p < 0.001$). On the Go/No-Go task, children with MLD made significantly more errors ($p < 0.001$), as reflected by their higher scores. Minute-based task-completion time differences between groups varied across tasks, with longer completion times most evident when tasks required multistep computation or mathematical reasoning, as in the Mental Calculation and Arithmetic Word Problem tasks. Composite-level internal-coherence analyses further support the conceptualization of Askisi-MD as a multidomain neuropsychological screener, rather than as a single homogeneous scale or a set of independent mathematics subscales. McDonald's Ω estimates indicated internal coherence for selected composites. Modest internal coherence was observed for the neuropsychological task-score composite and the total task-score composite. The strongest coherence was observed for the latency-based mathematical processing-efficiency composite. This finding suggests that minute-based completion-time indicators across mathematical tasks may capture a coherent dimension of task-completion efficiency. However, these results should not be interpreted as evidence that the brief mathematics tasks are reliable standalone psychometric subscales. Overall, Askisi-MD appears to provide internally coherent screening information at the composite level, while maintaining its role as a multidimensional first-line screener for identifying children who may require further neuropsychological and educational assessment.

Furthermore, the screener showed good case-control discrimination, balanced classification indices, and good cross-validated classification performance for the best-performing model. It is also important to note that validation was approached through two complementary analytic frameworks. Binary logistic regression was conducted in SPSS, consistent with conservative psychometric practice in neuropsychology, where the emphasis is on estimating how strongly each task contributes to the likelihood of group membership and on expressing this relationship through interpretable parameters. To support further predictive validation, additional supervised classification models were also employed. These models were not intended to test linear associations, as they evaluate whether multivariate patterns across tasks can reliably separate groups, even when the relationships that exist among variables are nonlinear or interactive. Their use in combination with stratified cross-validation places weight on performance stability across resampled folds, strengthening the argument that observed discrimination is not dependent on a single partition of the data.

In conclusion, Askisi-MD offers promise as a feasible first-line tool to screen for MLD in children, providing opportunities for comprehensive neuropsychological and educational assessment and early intervention when MLD may be present. However, external validation of our results and evaluation of the screener in real-world school settings remains essential.

Availability of Data and Materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

Conceptualization, NCZ, SKS, FV, PP, LM and GN; Data curation, NCZ and SKS; Formal analysis, NCZ, SKS, FV and ES; Investigation, NCZ; Methodology, NCZ, SKS, FV and ES; Interpretation of results, NCZ, SKS, FV, ES, PP, LM and GN; Visualization, NCZ, SKS and FV; Writing—original draft preparation, NCZ, SKS, FV and ES; Writing—review and editing, PP, LM and GN. All authors contributed to editorial changes in the manuscript. All authors have read and agreed to the published version of the manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

The study protocol was reviewed and approved by the Internal Research Ethics Committee of the Department of Informatics and Telecommunications, University of Thessaly (approval protocol code: 227092024). The study was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from the parents or legal guardians of all participating children before data collection.

Acknowledgment

The authors sincerely thank all participating children and their parents or legal guardians for their valuable time, cooperation, and contribution to this study.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/JIN52375.

References

[1] Butterworth B, Varma S, Laurillard D. Dyscalculia: from brain to education. Science (New York, N.Y.). 2011; 332: 1049–1053. https://doi.org/10.1126/science.1201536

[2] Chinn S. The trouble with maths: A practical guide to helping learners with numeracy difficulties. Routledge: UK. 2020.

[3] Manning C, Jennings B, MacLennan K, Rayat R, Spiller K, Joseph H. What Are the Research Priorities for the Dyslexia Community in the United Kingdom and How Do They Align With Previous Research Funding? Dyslexia (Chichester, England). 2025; 31: e70004. https://doi.org/10.1002/dys.70004

[4] Han W. Dyscalculia and dyslexia in school-aged children: comorbidity, support, and future prospects. Frontiers in education. Frontiers Media SA. 2025; 10: 1515216. https://doi.org/10.3389/feduc.2025.1515216

[5] Polydoros G. Investigating the Link Between Dyslexia and Dyscalculia: Teaching, Learning, and ICT Integration. Journal of Computer Assisted Learning. 2026; 42: e70233. https://doi.org/10.1002/jcal.70233

[6] Wilson AJ, Andrewes SG, Struthers H, Rowe VM, Bogdanovic R, Waldie KE. Dyscalculia and dyslexia in adults: Cognitive bases of comorbidity. Learning and Individual Differences. 2015; 37: 118–132. https://doi.org/10.1016/j.lindif.2014.11.017

[7] Luciano M, Ma AY, McKinlay M, Mubarak R, Paracchini S. Shaping UK research priorities in the area of specific learning difficulties by surveying people with lived experience. Neurodiversity. 2025; 3: 27546330251348559. https://doi.org/10.1177/27546330251348559

[8] Anderson KG. Gender bias and special education referrals. Annals of Dyslexia. 1997; 47: 151–162. https://doi.org/10.1007/s11881-997-0024-8

[9] Parsons S, Bynner J. Numeracy and employment. Education + Training. 1997; 39: 43–51. https://doi.org/10.1108/00400919710164125

[10] Parsons S, Bynner J. Does numeracy matter more? National Research and Development Centre for Adult Literacy and Numeracy. 2005.

[11] Karagiannakis G, Baccaglini-Frank A, Papadatos Y. Mathematical learning difficulties subtypes classification. Frontiers in Human Neuroscience. 2014; 8: 57. https://doi.org/10.3389/fnhum.2014.00057

[12] Li C, Wang L. Charting the knowledge landscape of mathematics learning disabilities: an analysis of intellectual structure and research frontiers. Frontiers in Psychology. 2025; 16: 1652056. https://doi.org/10.3389/fpsyg.2025.1652056

[13] Ulubaş SC, Gök M. Parental awareness, diagnostic processes, and support needs regarding mathematical learning difficulties. BMC Psychology. 2026; 14: 241. https://doi.org/10.1186/s40359-026-04025-3

[14] American Psychological Association (Ed.). Diagnostic and statistical manual of mental disorders: DSM-5. American Psychiatric Association Publishing: Washington, DC, USA. 2013.

[15] Cohen Kadosh R, Bahrami B, Walsh V, Butterworth B, Popescu T, Price CJ. Specialization in the human brain: the case of numbers. Frontiers in Human Neuroscience. 2011; 5: 62. https://doi.org/10.3389/fnhum.2011.00062

[16] Faye A, Jacquin-Courtois S, Reynaud E, Lesourd M, Besnard J, Osiurak F. Numerical cognition: A meta-analysis of neuroimaging, transcranial magnetic stimulation and brain-damaged patients studies. NeuroImage. Clinical. 2019; 24: 102053. https://doi.org/10.1016/j.nicl.2019.102053

[17] McCloskey M, Caramazza A, Basili A. Cognitive mechanisms in number processing and calculation: evidence from dyscalculia. Brain and Cognition. 1985; 4: 171–196. https://doi.org/10.1016/0278-2626(85)90069-7

[18] Dehaene S, Cohen L. Number processing. Mathematical cognition. 1996; 1: 83–120.

[19] Dehaene S. Précis of the number sense. Mind & Language, 2001; 16: 16–36. https://doi.org/10.1111/1468-0017.00154

[20] Dehaene S, Piazza M, Pinel P, Cohen L. Three parietal circuits for number processing. Cognitive neuropsychology. 2003; 20: 487–506. https://doi.org/10.1080/02643290244000239

[21] Jia L, Zhang Z, Wang C, Liu Y, Yu F, Zhang D, et al. A partially shared magnitude system: Common and specialized neural substrates underlying spatial, numerical, and temporal arithmetic. NeuroImage. 2026; 328: 121800. https://doi.org/10.1016/j.neuroimage.2026.121800

[22] Nieder A. The calculating brain. Physiological Reviews. 2025; 105: 267–314. https://doi.org/10.1152/physrev.00014.2024

[23] Khan OA, Peters L, Kwok FY, Lynn A, Ansari D, Wilkey ED. Neural connectivity in developmental dyscalculia: a resting-state fMRI study using ROI-based and connectome-based approaches. ZDM–Mathematics Education. 2026; 1–15.

[24] Michels L, O'Gorman R, Kucian K. Functional hyperconnectivity vanishes in children with developmental dyscalculia after numerical intervention. Developmental Cognitive Neuroscience. 2018; 30: 291–303. https://doi.org/10.1016/j.dcn.2017.03.005

[25] Michels L, Buechler R, Kucian K. Increased structural covariance in brain regions for number processing and memory in children with developmental dyscalculia. Journal of Neuroscience Research. 2022; 100: 522–536. https://doi.org/10.1002/jnr.24998

[26] Mateu-Estivill R, Adan A, Grau S, Rifà-Ros X, Caldú X, Bargalló N, et al. Alterations in functional brain connectivity associated with developmental dyscalculia. Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging. 2024; 34: 694–703. https://doi.org/10.1111/jon.13236

[27] Üstün S, Ayyıldız N, Kale EH, Mançe Çalışır Ö, Uran P, Öner Ö, et al. Children With Dyscalculia Show Hippocampal Hyperactivity During Symbolic Number Perception. Frontiers in Human Neuroscience. 2021; 15: 687476. https://doi.org/10.3389/fnhum.2021.687476

[28] Soltanlou M, Dresler T, Artemenko C, Rosenbaum D, Ehlis AC, Nuerk HC. Training causes activation increase in temporoparietal and parietal regions in children with mathematical disabilities. Brain Structure & Function. 2022; 227: 1757–1771. https://doi.org/10.1007/s00429-022-02470-5

[29] Schwizer Ashkenazi S, McCaskey U, O'Gorman Tuura R, Kucian K. Altered Effective Connectivity of the Numerical Brain in Children With Developmental Dyscalculia. Journal of Neuroscience Research. 2025; 103: e70066. https://doi.org/10.1002/jnr.70066

[30] von Aster MG, Shalev RS. Number development and developmental dyscalculia. Developmental Medicine and Child Neurology. 2007; 49: 868–873. https://doi.org/10.1111/j.1469-8749.2007.00868.x

[31] Kucian K, von Aster M. Developmental dyscalculia. European Journal of Pediatrics. 2015; 174: 1–13. https://doi.org/10.1007/s00431-014-2455-7

[32] Attout L, Majerus S. Working memory deficits in developmental dyscalculia: The importance of serial order. Child Neuropsychology: a Journal on Normal and Abnormal Development in

Childhood and Adolescence. 2015; 21: 432–450. https://doi.org/10.1080/09297049.2014.922170
[33] Starling-Alves I, Shanley L, Cook MA, Sabb FW, Smith J, Clarke B, et al. Children identified for classroom-based math support show altered resting-state connectivity in parietal brain regions. Scientific Reports. 2025; 15: 27013. https://doi.org/10.1038/s41598-025-12809-7
[34] Szucs D, Devine A, Soltesz F, Nobes A, Gabriel F. Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior. 2013; 49: 2674–2688. https://doi.org/10.1016/j.cortex.2013.06.007
[35] Zerouali S, Kaddouri H, El-Kamia A, Alaoui S. Direct effects of dyscalculia on executive functions: revisiting mediation models. Child Neuropsychology: a Journal on Normal and Abnormal Development in Childhood and Adolescence. 2026; 32: 730–748. https://doi.org/10.1080/09297049.2025.2610225
[36] Zhang J, Cheung SK, Wu C, Meng Y. Cognitive and Affective Correlates of Chinese Children's Mathematical Word Problem Solving. Frontiers in Psychology. 2018; 9: 2357. https://doi.org/10.3389/fpsyg.2018.02357
[37] Lai Y, Zhu X, Chen Y, Li Y. Effects of Mathematics Anxiety and Mathematical Metacognition on Word Problem Solving in Children with and without Mathematical Learning Difficulties. PloS One. 2015; 10: e0130570. https://doi.org/10.1371/journal.pone.0130570
[38] Agostini F, Zoccolotti P, Casagrande M. Domain-General Cognitive Skills in Children with Mathematical Difficulties and Dyscalculia: A Systematic Review of the Literature. Brain Sciences. 2022; 12: 239. https://doi.org/10.3390/brainsci12020239
[39] Wilkey ED, Pollack C, Price GR. Dyscalculia and Typical Math Achievement Are Associated With Individual Differences in Number-Specific Executive Function. Child Development. 2020; 91: 596–619. https://doi.org/10.1111/cdev.13194
[40] Kroesbergen EH, Huijsmans MDE, Friso-van den Bos I. A meta-analysis on the differences in mathematical and cognitive skills between individuals with and without mathematical learning disabilities. Review of Educational Research. 2023; 93: 718–755. https://doi.org/10.3102/00346543221132773
[41] Kuhn JT. Developmental dyscalculia. Zeitschrift für Psychologie. 2015.
[42] Mammarella IC, Hill F, Devine A, Caviola S, Szűcs D. Math anxiety and developmental dyscalculia: A study on working memory processes. Journal of Clinical and Experimental Neuropsychology. 2015; 37: 878–887. https://doi.org/10.1080/13803395.2015.1066759
[43] Yadav B, Yadav M. Breaking Down the Dyscalculia Barrier: The Subtle Role of Emotional Intelligence on Mathematical Performance. MIER Journal of Educational Studies Trends and Practices. 2025; 15: 464–489. https://doi.org/10.52634/mier/2025/v15/i2/2889
[44] Butterworth B. Dyscalculia Screener (software and manual). NFER-Nelson: London. 2003.
[45] Castaldi E, Tinelli F, Filippo G, Bartoli M, Anobile G. Auditory time perception impairment in children with developmental dyscalculia. Research in Developmental Disabilities. 2024; 149: 104733. https://doi.org/10.1016/j.ridd.2024.104733
[46] Zygouris NC, Stamoulis GI, Vlachos F, Vavougios D, Dadaliaris AN, Nerantzaki E, et al. Screening for Disorders of Mathematics via a web application', 2017 IEEE Global Engineering Education Conference (EDUCON). IEEE: Athens, Greece. 2017. https://doi.org/10.1109/EDUCON.2017.7942893
[47] Zygouris NC, Vlachos F, Dadaliaris AN, Oikonomou P, Stamoulis GI, Vavougios D, et al. A neuropsychological approach of developmental dyscalculia and a screening test via a web application. International Journal of Engineering Pedagogy (iJEP). 2017; 7: 51–65.
[48] Räsänen P, Laurillard, D, Käser T, von Aster M. Perspectives to technology-enhanced learning and teaching in mathematical learning difficulties. International Handbook of Mathematical Learning Difficulties: From the Laboratory to the Classroom (pp. 733–754). Springer International Publishing: Cham. 2019.
[49] Räsänen P, Aunio P, Laine A, Hakkarainen A, Väisänen E, Finell J, et al. Effects of Gender on Basic Numerical and Arithmetic Skills: Pilot Data From Third to Ninth Grade for a Large-Scale Online Dyscalculia Screener. Frontiers in Education. Frontiers Media SA. 2021; 6: 683672. https://doi.org/10.3389/feduc.2021.683672
[50] Tanchev M, Koceski S. Web-based educational game for early screening and support dyscalculia. Balkan Journal of Applied Mathematics and Informatics. 2025; 8: 43–54.
[51] Pudjoatmodjo B, Salam S, Pee ANC, Rudavan RA, Prihatmanto AS. Screening capabilities for the 3D dyscalculia identification game framework. Bulletin of Electrical Engineering and Informatics. 2025; 14: 3015–3028.
[52] Hewapathirana C, Abeysinghe K, Maheshani P, Liyanage P, Krishara J, Thelijjagoda S. A mobile-based screening and refinement system to identify the risk of dyscalculia and dysgraphia learning disabilities in primary school students', 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS). IEEE: Negambo, Sri Lanka. 2021. https://doi.org/10.1109/ICIAfS52090.2021.9605998
[53] Samarasinghe NR, Abeyasinghe AHMGB. Early diagnosis of the learning disabilities in kids using a computer game-based solution. Disability and Rehabilitation. Assistive Technology. 2023; 18: 1310–1318. https://doi.org/10.1080/17483107.2021.2003454
[54] Wilson AJ, Revkin SK, Cohen D, Cohen L, Dehaene S. An open trial assessment of "The Number Race", an adaptive computer game for remediation of dyscalculia. Behavioral and Brain Functions: BBF. 2006; 2: 20. https://doi.org/10.1186/1744-9081-2-20
[55] Butterworth B, Laurillard D. Low numeracy and dyscalculia: identification and intervention. Zdm. 2010; 42: 527–539.
[56] Miundy K, Zaman HB, Nordin A, Ng KH. Early intervention through identification of learners with dyscalculia as initial analysis to design AR assistive learning application. International Visual Informatics Conference (pp. 110–122). Springer International Publishing: Cham. 2019.
[57] Wen T, Yiqun GE. Mathematics, Information and Learning. Mathematics for Future Computing and Communications. 2021; 230.
[58] Zygouris NC, Toki EI, Vlachos F, Styliaras SK, Tziritas N. The Implementation of the Askisi-SD Neuropsychological Web-Based Screener: A Battery of Tasks for Screening Cognitive and Spelling Deficits of Children. Education Sciences. 2025; 15: 452. https://doi.org/10.3390/educsci15040452
[59] Zygouris NC, Vlachos F, Styliaras SK, Tziallas GD, Avramidis E. Validation of the Askisi-Lexia neuropsychological web-based screener: A neuropsychological battery for screening cognitive and phonological skills of children with dyslexia. Applied Neuropsychology. Child. 2025; 1–17. https://doi.org/10.1080/21622965.2025.2461192
[60] Georgas D, Paraskevopoulos I, Bezevegis E, Giannitsas N. Wechsler Intelligence Scale for Children-(WISC-III): Greek Version Manual. Ellinika Grammata: Athens. 1997.
[61] DuPaul GJ, Power TJ, Anastopoulos AD, Reid R. ADHD Rating Scale-IV: Checklists, norms, and clinical interpretation. Guilford Press: NY, USA. 1998.
[62] Kalantzi-Azizi A, Aggeli K Efstathiou G. Greek ADHD Rating Scale-IV: Standardization of the ADHD-IV Rating Scale by G. J. DuPaul, T. J. Power, A. D. Anastopoulos, and R. Reid. Ellinika

Grammata: Athens. 2006.
[63] Cangoz B, Altun A, Olkun S, Kacar F. Computer Based Screening Dyscalculia: Cognitive and Neuropsychological Correlates. Turkish Online Journal of Educational Technology-TOJET. 2013; 12: 33–38.
[64] Mazzocco MMM, Feigenson L, Halberda J. Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). Child Development. 2011; 82: 1224–1237. https://doi.org/10.1111/j.1467-8624.2011.01608.x
[65] Bugden S, Ansari D. Probing the nature of deficits in the 'Approximate Number System' in children with persistent Developmental Dyscalculia. Developmental Science. 2016; 19: 817–833. https://doi.org/10.1111/desc.12324
[66] Decarli G, Sella F, Lanfranchi S, Gerotto G, Gerola S, Cossu G, et al. Severe Developmental Dyscalculia Is Characterized by Core Deficits in Both Symbolic and Nonsymbolic Number Sense. Psychological Science. 2023; 34: 8–21. https://doi.org/10.1177/09567976221097947
[67] Lamb S, Krieger F, Kuhn JT. Delayed development of basic numerical skills in children with developmental dyscalculia. Frontiers in Psychology. 2024; 14: 1187785. https://doi.org/10.3389/fpsyg.2023.1187785
[68] Abreu-Mendoza RA, Chamorro Y, Garcia-Barrera MA, Matute E. The contributions of executive functions to mathematical learning difficulties and mathematical talent during adolescence. PloS One. 2018; 13: e0209267. https://doi.org/10.1371/journal.pone.0209267
[69] Peng P, Fuchs D. A meta-analysis of working memory deficits in children with learning difficulties: Is there a difference between verbal domain and numerical domain? Journal of Learning Disabilities. 2016; 49: 3–20. https://doi.org/10.1177/0022219414521667
[70] Fuchs L, Fuchs D, Seethaler PM, Barnes MA. Addressing the role of working memory in mathematical word-problem solving when designing intervention for struggling learners. ZDM Mathematics Education. 2020; 52: 87–96. https://doi.org/10.1007/s11858-019-01070-8
[71] Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology (London, England: 1953). 2014; 105: 399–412. https://doi.org/10.1111/bjop.12046
[72] Amland T, Grande G, Scherer R, Lervåg A, Melby-Lervåg M. Cognitive factors underlying mathematical skills: A systematic review and meta-analysis. Psychological Bulletin. 2025; 151: 88–129. https://doi.org/10.1037/bul0000457
[73] Geary DC, Hoard MK, Nugent L, Bailey DH. Mathematical Cognition Deficits in Children With Learning Disabilities and Persistent Low Achievement: A Five-Year Prospective Study. Journal of Educational Psychology. 2012; 104: 206–223. https://doi.org/10.1037/a0025398
[74] Peng P, Wang C, Namkung J. Understanding the cognition related to mathematics difficulties: A meta-analysis on the cognitive deficit profiles and the bottleneck theory. Review of Educational Research. 2018; 88: 434–476. https://doi.org/10.3102/0034654317753350
[75] Menon V. Working memory in children's math learning and its disruption in dyscalculia. Current Opinion in Behavioral Sciences. 2016; 10: 125–132. https://doi.org/10.1016/j.cobeha.2016.05.014
[76] Raghubar KP, Barnes MA, Hecht SA. Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. Learning and Individual Differences. 2010; 20: 110–122. https://doi.org/10.1016/j.lindif.2009.10.005
[77] Winkel K, Zipperle I. Children with mathematical learning difficulties—How do their working memory skills differ from typically developing first graders? Journal für Mathematik-Didaktik. 2023; 44: 417–440.
[78] Rodrigues J, Jordan NC, Hansen N. Identifying Fraction Measures as Screeners of Mathematics Risk Status. Journal of Learning Disabilities. 2019; 52: 480–497. https://doi.org/10.1177/0022219419879684
[79] Martin BN, Fuchs LS. Predicting Risk for Comorbid Reading and Mathematics Disability Using Fluency-Based Screening Assessments. Learning Disabilities Research & Practice: a Publication of the Division for Learning Disabilities, Council for Exceptional Children. 2022; 37: 100–112. https://doi.org/10.1111/ldrp.12278
[80] de León SC, Jiménez JE, García E, Gutiérrez N, Gil V. Universal screening in mathematics for Spanish students in first grade. Learning Disability Quarterly. 2021; 44: 123–135. https://doi.org/10.1177/0731948720903273
[81] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. 1995; 14: 1137–1145.
[82] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794. https://doi.org/10.1145/2939672.2939785
[83] Breiman L. Random Forests. Machine Learning. 2001; 45: 5–32. https://doi.org/10.1023/A:1010933404324
[84] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27: 861–874.
[85] Rodríguez C, Jiménez JE, de León SC, Marco I. Multimedia Battery for Assessment of Cognitive and Basic Skills in Mathematics (BM-PROMA). Journal of Visualized Experiments: JoVE. 2021; 10.3791/62288. https://doi.org/10.3791/62288
[86] Kariyawasam R, Nadeeshani M, Hamid T, Subasinghe I, Ratnayake P. A gamified approach for screening and intervention of dyslexia, dysgraphia and dyscalculia', 2019 international conference on advancements in computing (icac). IEEE: Malabe, Sri Lanka. 2019.
[87] Bhushan S, Arunkumar S, Eisa TAE, Nasser M, Singh AK, Kumar P. AI-Enhanced Dyscalculia Screening: A Survey of Methods and Applications for Children. Diagnostics (Basel, Switzerland). 2024; 14: 1441. https://doi.org/10.3390/diagnostics14131441
[88] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017; 30.
[89] Lariviere DO, Powell SR, Fall AM, Roberts G, Arsenault TL. Language Predictors of Word-Problem Performance Among Third-Grade Students With Mathematics Difficulty. Journal of Learning Disabilities. 2025; 58: 445–458. https://doi.org/10.1177/00222194241311979
[90] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ (Clinical Research Ed.). 2024; 385: e078378. https://doi.org/10.1136/bmj-2023-078378
[91] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clinical Kidney Journal. 2020; 14: 49–58. https://doi.org/10.1093/ckj/sfaa188
[92] Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Medicine. 2023; 21: 70. https://doi.org/10.1186/s12916-023-02779-w