Artificial Intelligence Item Analysis Tool for Educational Assessment: Case of Large-Scale Competitive Exams

Najoua Hrich[1,*], Mohamed Azekri[2], and Mohamed Khaldi[3]

[1]Regional Center of Education & Training Professions, Institutions for Higher Executive Training, Tangier, Morocco
[2]Regional Académie of Education & Training, Ministry of National Education Preschool and Sports, Tetouan, Morocco
[3]Higher Normal School, Abdelmalek Essaadi University, Tetouan, Morocco
Email: hrnajouaofficiel@gmail.com (N.H.); medazekri@gmail.com (M.A.); medkhaldi@yahoo.fr (M.K.)
[*]Corresponding author
Manuscript received December 10, 2023; revised December 25, 2023; accepted February 1, 2024; published June 17, 2024

***Abstract*—With the increased number of competitive examinees, adopting Multiple Choice Tests (MCTs) in most examinations has significantly shaped the assessment methodology. However, the success of this method depends on the quality of the items. Thus, selecting relevant items, balanced for difficulty and discrimination power, is crucial to guarantee the assessments' validity and reliability. In this regard, integrating Artificial Intelligence (AI) provides promising prospects for further enhancing the item analysis and selection process. Therefore, this research aims to build a Machine-Learning (ML) model that discerns and selects items based on their difficulty and discrimination. This study employs the Artificial Neural Networks (ANN) method through binary classification models for item classification. The study's experimental results demonstrate the proposed model's efficacy, showcasing superior performance with an accuracy rate of 96% for item selection.**

***Keywords*—e-assessment, competitive exams, items analysis, P-index, D-index, artificial intelligence, deep learning**

I. Introduction

In recent years, the Ministry of National Education in Morocco has digitized its services, progressively moving away from traditional paper-based methods to digital procedures. This evolution is guided by technological advancements and the widespread of Artificial Intelligence. Simultaneously, the Ministry has significantly changed its recruitment methodology, adopting Multiple Choice Tests (MCTs) for all competitive exams. This shift is part of an initiative to modernize the selection processes for educational personnel. MCTs offer several advantages, including a more objective assessment of candidates' skills and knowledge. However, considering the items' difficulty level and discriminative power is crucial to this transition. The precise adjustment of these parameters directly influences the selection of candidates. In this context, item analysis is an essential step in test development. It ensures that the items are valid, reliable, and exhibit high discrimination power [1]. Habitually, this process is led by experts who carefully design and analyze each question based on their knowledge and expertise. However, this approach can be laborious and subjective, leading to bias and inconsistency. AI and ML technologies have emerged as powerful tools for facilitating the analysis of items in competitive exams. By using Artificial Intelligence (AI) and Machine-Learning (ML) [2], exam administrators can improve the reliability and fairness of exams by selecting items that are effective and free from bias. By automating specific tasks involved in item analysis, AI can help reduce the time and effort required for item selection; this can be particularly valuable for large-scale competitive exams with numerous items [3].

The contribution of this paper is to propose a model combining statistical techniques basis and Deep Learning (DL) that allows the selection of practical items based on their difficulty and discrimination indexes. Regenerating results based on candidates' performances by eliminating scores of inappropriate items. Additionally, it regenerates candidates' scores after removing inappropriate items.

To validate this model, we collected data from 3600 participants in the recruitment competitions for future computer science teachers administered by the Ministry of National Education Preschool and Sports for the 2022 and 2023 academic years. Each year's competition tests, comprising 120 single-choice items, underwent rigorous analysis involving the calculation of difficulty and discrimination indexes. Subsequently, we developed an artificial neural network to discern and select the most discriminating items to regenerate candidates' scores.

II. Related Works

A. Large-Scale Competitive Exams

Large-scale competitions assess the knowledge and skills of candidates wishing to access educational establishments or employment. These tests are usually administered at the national or regional level and are very competitive, with many applicants competing for a limited number of seats and positions. To select the most qualified individuals, this type of exam must permit candidates to be differentiated based on where they stand on the evaluated dimension. Competitive exams can include a variety of subjects, such as mathematics, computer sciences, languages, and social studies. They may be conducted online or offline, ranging from objective multiple-choice tests to subjective essay-based items.

When an assessment process aims to differentiate individuals according to a given criterion (their level of competence, mastery, attitude, motivation, etc.), we must use items that have a high power of discrimination (the ability to distinguish as clearly and as finely as possible individuals according to the considered criterion). MCTs are the most commonly used formats in large-scale competitive exams, whether in person or remotely. One characteristic of the MCTs is that candidates can answer correctly by chance. For example, in a multiple-choice test with two options, where only one is correct, the chance of correctly answering is 50%. It is crucial to consider the effect of these random responses [4]. Therefore, conducting a detailed analysis of

candidates' responses becomes essential. The following section will explore this issue further, presenting various statistics to assess the quality of the test items [5].

B. Item Analysis

Item analysis is a statistical technique used to evaluate the quality of test items in educational assessment. It involves analyzing the responses of students to each item to identify items that are too easy or too difficult, items that are not discriminating enough (i.e., don't differentiate between high-performing and low-performing students), and items with high rates of guessing. In item analysis [6], several statistics are commonly used to evaluate the quality of test items, which include:

- Item difficulty: the proportion of students who answered the item correctly. Items with very high or very low difficulty may be problematic.
- Item discrimination: the degree to which an item differentiates between high-performing and low-performing students. High discrimination is desirable, indicating that the item measures the intended construct.
- Item-total correlation: the correlation between an item and the total test score. This indicates how well the item measures the same construct as the rest of the test.
- Point-biserial correlation: the correlation between an item and the total test score, considering whether the student answered the item correctly or incorrectly. This is used for dichotomous (true/false or multiple-choice) items.
- Distractor analysis: an analysis of the responses to each distractor (incorrect option) in a multiple-choice item to identify which distractors were most chosen and whether any are highly correlated with the correct answer.

Several tools have emerged to facilitate this analysis, each offering specific functionalities to meet the varied needs of researchers and practitioners. Table 1 presents some widely used tools.

Although these tools are useful for test analysis, they have limitations regarding assessment customization and item selection. Predefined Item Response Theory (IRT) models may restrict the Classical Item and Test Analysis Spreadsheet (CITAS) Platform; Excel lacks advanced analysis based on item response theory, and JMetrik may require advanced skills.

Table 1. Analysis items tools

Tool	Description	What can do
Classical Item and Test Analysis Spreadsheet (CITAS) [7]	It is an easy-to-use tool for implementing classical test theory on small data sets, designed to provide a straightforward and no-cost way for non-psychometricians to evaluate the quality of assessments.	- Mean: The average score. - Standard deviation: An index of the variation in scores. - Reliability: An index of test quality on a scale of 0 to 1, using coefficient alpha (aka KR20). - Standard Error of Measurement (SEM): An index of score error that can be used to create confidence intervals with a classical test theory approach. - Item P values: An item difficulty statistic. - Item point-biserial: An item discrimination statistic. - Distractor analysis: Frequencies of each item response.
Excel for Classic Test analysis [8]	Calculate the basic statistics developed in classical test analysis for closed response items such as multiple choice.	- Distractor analysis - Item facility - Discrimination index - Reliability - Descriptive statistics (mean, standard deviation, and standard measurement error). - Cronbach's alpha
JMetrik [9]	It is free and open-source psychometric software.	Psychometric methods include classical item analysis, reliability estimation, test scaling, differential item functioning.

C. Artificial Intelligence for Assessment

AI is increasingly used in various assessment aspects, from item analysis and selection to automated scoring and feedback generation. AI can improve the efficiency and fairness of assessments by reducing human bias, enhancing objectivity, and providing personalized learning experiences [10]. Table 2 presents some applications of AI in educational assessment.

Table 2. Applications of AI in educational assessment

Feature	Description
Automated scoring	AI automates the grading of specific assessments, like multiple-choice tests or essays. Through analyzing the language and content of a submission, AI algorithms assess the structure, grammar, and other elements of a student's writing, ultimately saving time and enhancing the fairness and consistency of evaluation [11].
Feedback generation	refers to providing feedback to students based on their assessment performance. AI can generate personalized feedback for each student, which can help them identify their strengths and challenges and subsequently improve their learning outcomes [12].
Personalized learning	AI is also used to personalize learning and assessment experiences. AI algorithms analyze students' data to identify strengths and weaknesses and recommend learning resources or assessment tasks that target improvement areas [13].
Cognitive diagnosis	using AI algorithms to analyze student responses to assessment items and diagnose their cognitive strengths and weaknesses. This approach goes beyond traditional IRT models, which focus on measuring overall ability or proficiency in a particular subject [14].
Cheating detection	AI is employed to identify cheating in e-assessments, including analyzing examinee behavior and recognizing cheating patterns, such as identical responses or abnormal response times [15].
Test security	AI enhances the security of the assessment by detecting and preventing fraud, such as impersonation, hacking, or distribution of exam materials [16].
Item analysis and selection	AI algorithms assist in this process by analyzing large amounts of data and identifying patterns and relationships that may not be immediately apparent to human evaluators. One example of using AI in analyzing items is using ML algorithms to identify which test items are most effective in differentiating between high-performing and low-performing students by helping to identify too easy or complex items, remove them from future assessments, and determine which items are most effective at measuring specific learning outcomes. AI can also be used in item selection, which involves selecting a set of items from a more significant item pool to create an assessment test. AI can assist in this process by selecting items that are well-matched to the intended learning outcomes and that maximize the information obtained from the assessment; this can help improve the validity and reliability of the evaluation and provide a more accurate image of student learning [17].

Among the most powerful AI tools for item selection, Artificial Neural Networks (ANNs) are prominent. The ANNs can model complex relationships between candidate performance and item characteristics, allowing dynamic personalization of tests. Additionally, ANNs can identify non-linear patterns in data, thereby improving the accuracy of item selection compared to traditional approaches [18].

III. Proposed Approach

In the context of this research, we propose an innovative approach that aims to significantly improve the selection of items in the field of evaluation based on statistical techniques to calculate difficulty and discrimination indexes and ANN to select appropriate items. The process of our proposal is shown in Fig. 1.

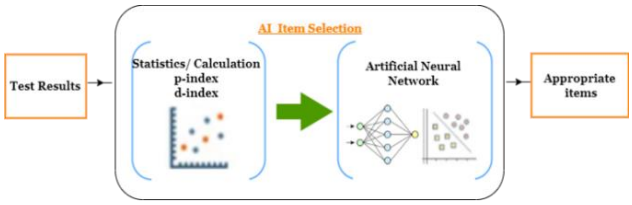

Fig. 1. AI Selection Item Process.

We have exclusively focused on item difficulty, expressed by the power index (P-index), and item discrimination, expressed by the discrimination index (D-index) [5].

- P-index: calculates the difficulty index for each item. The difficulty index is the ratio of test-takers who answered the item correctly.

$$\text{P-index} = \frac{\text{Number of correct responses}}{\text{Total number of respondents}}$$

- D-index: measures how well an item differentiates between high and low performers. It is typically computed by comparing the performance of the top group (e.g., the top 27% of scores) with the bottom group (e.g., the bottom 27% of scores) on the overall test.

$$\text{D-index} = \frac{\text{Top Group Mean} - \text{Bottom Group Mean}}{\text{Standard Deviation of total Scores}}$$

These indexes provide an in-depth understanding of an item's ability to discriminate among candidate performances and its inherent complexity.

Concurrently, we leverage the advantages of ANN in the item selection process. ANNs, with their input, hidden, and output layers can learn complex representations from item features. Their ability to capture non-linear patterns and model complex relationships between difficulty, discrimination, and item quality is particularly relevant in this context. ANN offers increased adaptability and flexibility to handle heterogeneous data, contributing to a more refined and accurate item selection.

This hybrid approach capitalizes on the robustness of traditional indexes while harnessing the power of deep learning provided by ANN. Our work will delve into the detailed implementation of this approach, highlighting how these two components complement each other to create a comprehensive and effective methodology in the assessment field. We will specifically emphasize the synergy between classical statistical analysis and the advanced modeling capabilities of ANN, underscoring a significant contribution to improving item selection processes (Fig. 2).

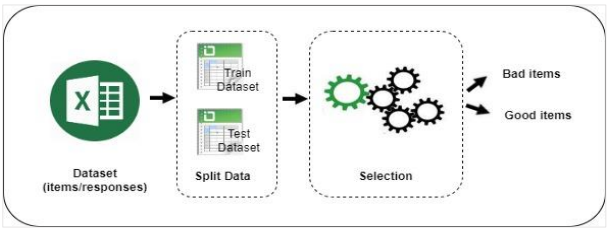

Fig. 2. Overview of the Deep Learning Selection Process.

IV. Methods and Materials

This research's methodology involves statistical techniques and an ANN DL model. We first calculated the difficulty and discrimination indexes using the results of recruitment competitions for future teachers in Morocco. Then, we developed an artificial neural network to select the most discriminating items to regenerate the candidates' scores.

The general steps are illustrated in the model depicted in Fig. 3.

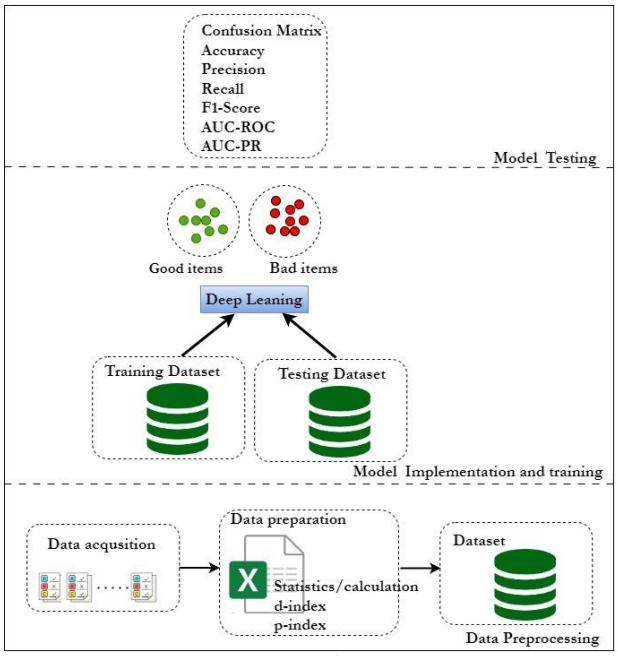

Fig. 3. Outline of the proposal.

The following sections cover a detailed description of each phase.

A. Data Preprocessing

The dataset was collected from the responses of 3,600 participants (for two academic years: 2022 and 2023) on the recruitment competitions for future computer science teachers, competitions administered by the Ministry of National Education, Preschool and Sports. Each year, the tests for the competition contain 120 single-choice items.

We have recorded the results in an Excel file containing two sheets, each presenting the data for a year. The columns represent items, and the lines represent candidates; their intersection indicates the candidate's score in an item (1 for the true answer and 0 for the wrong one). Fig. 4 presents an extract of our dataset.

Items	Id_C1	Id_C2	Id_C3	Id_C4	Id_C5	Id_C6	Id_C7	Id_C8	Id_C9
Item1	0	0	1	1	1	0	1	0	1
Item2	0	0	1	1	0	0	1	0	1
Item3	1	0	1	1	1	0	0	1	0
Item4	0	0	0	0	0	0	0	1	0
Item5	0	0	0	0	1	0	0	0	1
Item6	1	1	0	0	0	1	0	1	0
Item7	1	0	0	1	0	0	0	1	0

Fig. 4. Dataset extract.

We subsequently developed a Python script specialized in statistical analysis using the Numpy and Pandas libraries. This script allowed the precise calculation of the P-index and d-index. For the training phase, Fig. 5 presents the results obtained. And Fig. 6 presents an extract of the results obtained for the testing phase.

Items	p_index	d_index	Selection
Item1	0.50	0.19	1
Item2	0.48	0.18	1
Item3	0.52	0.02	0
Item4	0.52	0.06	0
Item5	0.50	0.13	1
...	...	...	...
Item116	0.56	0.10	0
Item117	0.49	0.12	1
Item118	0.54	0.11	1
Item119	1.00	0.01	0
Item120	0.55	0.18	1

Fig. 5. Extract of training dataset.

Items	p_index	d_index
Item1	0.53	-0.03
Item2	0.47	-0.02
Item3	0.53	0.00
Item4	0.48	-0.07
Item5	0.47	0.05
...	...	...
Item116	0.50	-0.01
Item117	0.51	-0.07
Item118	0.48	-0.07
Item119	0.49	0.05
Item120	0.47	-0.01

Fig. 6. Extract of testing dataset.

B. Model Architecture

DL models are machine learning models that map a set of predictor variables through a sequence of transformations called layers to predict a set of outcome variables. Much of DL's success in recent years can be attributed to a family of nonlinear statistical models called ANNs [19].

The type of neural network used in our model is a standard ANN. It is a dense feedforward neural network called Multi-layer Perception (MLP) (Fig. 7).

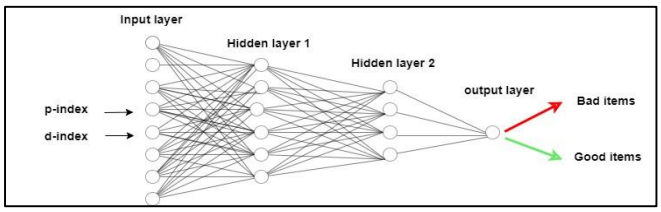

Fig. 7. Deep Neural Networks Architecture.

The model comprises four layers, including an input layer with 256 neurons that introduces nonlinearity through a Rectifier Linear Unit (ReLU) activation function commonly employed in neural networks. Two hidden layers follow, with 128 and 64 neurons utilizing the "ReLU" activation function. The final layer is the output layer, featuring one neuron and employing a sigmoid activation function suitable for binary classification. The output of this layer represents a probability ranging from 0 to 1.

C. Model Implementation

The module is developed under Python, the most used and famous programming language in data science. Python is a high-level programming language, and its basic design philosophy is based on code readability and syntax that allows programmers to express concepts in just a few lines of code. Python is an open-source license, making it freely usable [20].

The implementation of the model occurs in two stages. Firstly, a Python script is developed for statistical analysis using the NumPy and Pandas libraries for data preprocessing. This script is the initial phase of model development. Fig. 8 displays an excerpt from the code demonstrating the calculations for P-index and D-index.

```
def calculate_p_index(row):
    row_numeric = pd.to_numeric(row[1:], errors='coerce')
    total_correct_responses = row_numeric.sum()
    total_respondents = len(row_numeric)
    p_index = round(total_correct_responses / total_respondents,2)
    return p_index
def calculate_d_index(row):
    pindexh = row['ph_index']
    pindexl = row['pl_index']
    d_indexval = pindexh - pindexl
```

Fig. 8. Extract of P-index and D-index calculation.

Subsequently, the model is defined using the TensorFlow sequential Application Programming Interface (API). It has four Dense Layers (fully connected). The first layer has 256 neurons with ReLU activation function; the two-second layers have, respectively, 128 and 64 neurons with ReLU activation function, and the last one has one neuron with Sigmoid activation, suitable for binary classification. The Adam optimizer is used with a binary loss function (binary cross-entropy) for model compilation, and the accuracy metric is also specified. The model is trained on the training data for 100 epochs with a batch size 32.

Fig. 9 shows a section of code that illustrates model creation.

```
#Model Definition
model = tf.keras.Sequential([
    tf.keras.layers.Dense(256, activation='relu', input_shape=(2,)),
    tf.keras.layers.Dense(128, activation='relu'),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

#Model Compilation
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

Fig. 9. Creation model.

D. Model Experiment

Testing the model on a real dataset is essential to evaluate its effectiveness. During this experimental phase, the model will be validated to determine its ability to accurately identify relevant items based on their difficulty level and discrimination power. The obtained results will provide crucial information on the performance of the model and its practical relevance in real-life situations.

As part of model testing, we employ the data from the first Excel sheet for training, providing essential information to

the model. Following that, we reserve the data from the second Excel sheet for the testing phase, ensuring an independent evaluation of the model's performance on unseen data.

The activation function selected is the "sigmoid" function, and the loss function used is the binary cross entropy. The Adam optimizer is chosen with a learning rate of 0.001. This sequential methodology allows a structured and efficient implementation of the model.

E. Model Evaluation

In the context of neural networks, several metrics are used to evaluate the performance of a model. We have used the most frequently encountered ones, named in Table 3 [21]:

Table 3. Evaluation metrics

Metric	Description	Value
Accuracy	This is the ratio of correct predictions to all predictions. It is a general metric that measures the model's ability to predict all classes correctly.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	It measures the precision of optimistic predictions. It is the ratio of true positives to all optimistic predictions.	$\frac{TP}{TP + FP}$
Recall (sensitivity)	It measures the model's capacity to find all occurrences of the positive class. It is the ratio of true positives to all actual occurrences of the positive class.	$\frac{TP}{TP + FN}$
F1-Score	This is the harmonic average of precision and recall. It is often used when classes are unbalanced.	$\frac{2 \times precision \times recal}{precision + recall}$
Area under the Receiver Operating Characteristic Curve (AUC-ROC)	It Measures the model's ability to discriminate between classes. A value of 1.0 indicates perfect discrimination.	
Area under the Precision-Recall Curve (AUC-PRC)	It Measures the precision of the model on positive examples. Like AUC-ROC, a value of 1.0 indicates perfect performance.	

These metrics are calculated using Scikit-learn libraries to evaluate the performance of our neural network model. The results obtained are presented in the next section.

V. Proposal Validation and Results

The evaluation of the model is based on several key metrics that provide an in-depth understanding of its performance. The confusion matrix gives a detailed overview of correct and incorrect predictions, illustrating the model's ability to discern between good and wrong items.

To obtain these metrics, a script developed in Python and exploiting the "sklearn. Metrics" library provided the following results (Figs. 10 and 11).

An in-depth analysis of those results confirms the exceptional performance of the item selection model based on difficulty and discrimination indexes.

The confusion matrix reveals that out of 540 items in class 0 (wrong items), the model correctly predicted 516 items, while for class 1 (good items), 612 items were correctly predicted.

```
Confusion Matrix:
[[516  24]
 [ 24 612]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       540
           1       0.96      0.96      0.96       636

    accuracy                           0.96      1176
   macro avg       0.96      0.96      0.96      1176
weighted avg       0.96      0.96      0.96      1176


Accuracy: 0.9591836734693877

Area under the ROC curve (AUC-ROC): 0.958909853249476

Area under the Precision-Recall curve (AUC-PR): 0.9724682325760493
```

Fig. 10. Confusion matrix and classification report.

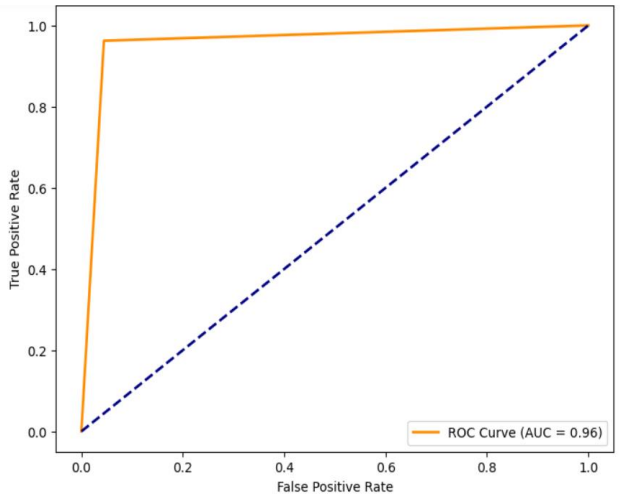

Fig. 11. ROC curve: Graphical representation that illustrates the performance of our binary classification model.

The classification report provides detailed information on precision, recall, and F1-Score for each class. The high values of these measures indicate that the model could differentiate well between difficult and discriminative items and those that were not. The overall accuracy of 95.92% confirms the model's reliability across the entire dataset.

The Area under the Receiver Operating Characteristic Curve (AUC-ROC) and the Area under the Precision-Recall Curve (AUC-PRC) are crucial metrics for evaluating the model's generalization ability. With values around 96%, the model maintains excellent performance on different datasets, suggesting its robustness and adaptability.

VI. Discussion

The results of this experiment are of particular importance in the context of large-scale competitions, as illustrated by the recruitment of future teachers in Morocco. The success of this approach demonstrates its relevance and practical applicability in extensive selection processes, where precision in candidate evaluation is crucial.

The joint integration of statistical techniques and a DL model in the item analysis provides significant advantages. DL models enable the precise selection of appropriate items, while statistical techniques provide rigor for calculating different indexes. In addition, this combination offers the possibility of personalizing the evaluation by excluding inappropriate items based on the characteristics of each item.

However, it is essential to note that further experience,

training, and experimentation with a more diverse and extensive dataset are necessary to enhance the robustness and generalizability of our proposed model. In conclusion, the results highlight the proposed model's remarkable effectiveness in item selection, characterized by its precision, sensitivity, and ability to generalize other data. These findings reinforce the model's validity and relevance in practical evaluation contexts.

VII. Conclusion

In conclusion, integrating DL approaches and statistical analysis in administering large-scale competitive exams can significantly improve the validity and reliability of the tests and results. By using the proposal model to analyze the test items and detect potential issues, educators and test developers can make data-driven decisions to improve the quality of the assessment. Additionally, they can gain deeper insights into candidates 'performance and make more informed decisions about instruction.

This proposed model is a component of a larger initiative where it will be improved by incorporating other analytical indexes. The perspective is to implement it on a specialized platform for building an item database, subsequently used for generating items for different exams.

Overall, the perspective of extending the experience of our model with other analytical indexes presents an exciting opportunity for improving the quality and validity of assessments. With continued research and development, this approach will potentially transform the field of education and provide more accurate and meaningful information about students' learning and performance.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

NH and MA conceptualized, prepared the research background, and developed the model; MK verified methods and supervised the findings of this work. All authors approved the final version.

References

[1] A. Rezigalla, "Item analysis: Concept and application," *Medical Education for the 21st Century*, 2022, pp. 1–17. doi: 10.5772/intechopen.100138

[2] Y. Liu and M. Baucham, "AI technology: Key to successful assessment," in *Handbook of Research on Redesigning Teaching, Learning, and Assessment in the Digital Era*, IGI Global, 2023, pp. 304–325. doi: 10.4018/978-1-6684-8292-6.ch016

[3] X. Zhai *et al.*, "A review of Artificial Intelligence (AI) in education from 2010 to 2020," *Complexity*, vol. 2021, pp. 1–18, 2021. doi: 10.1155/2021/8812542

[4] N. Hrich, M. Lazaar, and M. Khaldi, "Improving Cognitive decision-making into adaptive educational systems through a diagnosis tool based on the competency approach," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 07, 226, 2019. doi: 10.3991/ijet.v14i07.9870

[5] SCALP. Quotation of MCQs and VF. [Online]. Available: https://www.psychometrie.jlroulin.fr/cours/aide_quizz.html?E33.html (in French)

[6] T. Moses, "A review of developments and applications in item analysis," in *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS*, R. E. Bennett and M. von Davier, Éds. Cham: Springer International Publishing, 2017, pp. 19–46. doi: 10.1007/978-3-319-58689-2_2

[7] CITAS: Classical Item & Test Analysis Spreadsheet CITAS. Assessment Systems. [Online]. Available: https://assess.com/citas/

[8] Excel Spreadsheets for Classical Test Analysis. [Online]. Available: https://languagetesting.info/statistics/excel.html

[9] IRT Illustrator. Psychomeasurement Systems. [Online]. Available: https://itemanalysis.com/irt-illustrator/

[10] N. Hrich, M. Azekri, and M. Khaldi, "Artificial intelligence for educational assessment," in *Proc. the 16th Annual International Conference of Education, Research and Innovation*, Seville, Spain, 2023, pp. 2120–2124. doi: 10.21125/iceri.2023.0598

[11] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, vol. 2, no. 2, 100050, 2023. doi: 10.1016/j.rmal.2023.100050

[12] A. V. Y. Lee, "Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation," *Studies in Educational Evaluation*, vol. 77, 101250, 2023. doi: 10.1016/j.stueduc.2023.101250

[13] A. Y. Q. Huang, O. H. T. Lu, and S. J. H. Yang, "Effects of artificial Intelligence–Enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom," *Computers & Education*, vol. 194, 104684, 2023. doi: 10.1016/j.compedu.2022.104684

[14] Z. Swiecki *et al.*, "Assessment in the age of artificial intelligence," *Computers and Education: Artificial Intelligence*, vol. 3, 100075, 2022. doi: 10.1016/j.caeai.2022.100075

[15] S. Kaddoura and A. Gumaei, "Towards effective and efficient online exam systems using deep learning-based cheating detection approach," *Intelligent Systems with Applications*, vol. 16, 200153, 2022. doi: 10.1016/j.iswa.2022.200153

[16] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *Int J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–204, 2020. doi: 10.1007/s40593-019-00186-y

[17] F. Martínez-Plumed, R. B. C. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo, "Item response theory in AI: Analysing machine learning classifiers at the instance level," *Artificial Intelligence*, vol. 271, pp. 18–42, 2019. doi: 10.1016/j.artint.2018.09.004

[18] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019. doi: 10.1109/ACCESS.2019.2912200

[19] C. J. Urban and D. J. Bauer, "A deep learning algorithm for high-dimensional exploratory item factor analysis," arXiv preprint, arxiv.2001.07859, 2021.

[20] F. Chollet, *Deep Learning with Python*, Manning Publications, 2017.

[21] Evaluate your ML.NET model with metrics. [Online]. Available: https://learn.microsoft.com/fr-fr/dotnet/machine-learning/resources/metrics (in French)