A Gamified Web Platform for the Automated Diagnosis of Childhood Phonological and Phonetic Disorders through Deep Learning

Josty Gerardo Tafur-Gonzales
Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
u2020c069@upc.edu.pe

Joao Arturo Basauri-Bazalar
Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
u201716123@upc.edu.pe

Sandra Wong-Durand
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
pcsiswon@upc.edu.pe (corresponding author)

Pedro Castaneda
Faculty of Systems Engineering and Electrical Mechanics, Universidad Nacional Toribio Rodriguez de Mendoza, Amazonas, Peru
pedro.castaneda@untrm.edu.pe

Alejandra Onate-Andino
Escuela Superior Politecnica de Chimborazo (ESPOCH), Riobamba, Ecuador
monate@espoch.edu.ec

Received: 10 December 2025 | Revised: 21 January 2026 and 8 February 2026 | Accepted: 19 February 2026

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.16859*

ABSTRACT

This paper presents a gamified web platform for the automated diagnosis of children's phonetic–phonological disorders. The system integrates deep learning models with acoustic representations extracted using Wav2Vec2 and structured linguistic coding. It was evaluated on a clinical corpus of over 700 recordings, using cross-validation and a comparison between seven classification models. The model based on deep dense networks achieved an accuracy of 83.57%, exceeding the commonly accepted clinical threshold. In addition, the system reduced the evaluation time by 49.6% compared to the traditional method. The system was preliminarily evaluated using speech data collected from 10 children, focusing on technical feasibility and performance trends rather than definitive clinical validation. While the obtained results show promising classification accuracy, they should be interpreted as an initial proof of concept. The results support its applicability as an objective, accessible, and scalable tool in clinical and educational contexts.

Keywords-speech sound disorders; deep learning; diagnostic automation; pediatric speech therapy; Wav2Vec 2.0; gamified platform; Spanish language processing; web-based evaluation tools

I. INTRODUCTION

Phonetic–phonological disorders, or speech sound disorders, involve difficulties in accurately producing speech sounds, ranging from isolated articulation errors to systematic phonological patterns that reduce intelligibility [1, 2]. These disorders affect 4–10% of children and, when persistent, interfere with language development and are associated with later difficulties in reading and writing, as well as challenges in social and academic interactions. Delayed or absent intervention may lead to educational, emotional, and social consequences, generating long-term impacts on the child and

broader society [2, 3]. Given these implications, improving diagnostic processes for speech disorders is clinically and socially relevant [4], particularly in a context where communication impairments limit participation and academic performance, and where a global shortage of speech-language pathologists—70% reporting waiting lists—delays prompt care [5].

In response to these difficulties, artificial-intelligence-based digital solutions have gained attention for supporting early and objective assessment. Neural networks trained on children's speech corpora have achieved accuracies above 70% in multi-category disorder classification tasks across diverse languages [6, 7], whereas fast-screening frameworks using deep learning have surpassed 90% accuracy [8], and automatic measures of "pronunciation quality" have replicated expert judgments in evaluating infant phonemes [3]. In parallel, web and mobile therapy platforms have improved access to clinical resources and strengthened engagement—particularly for young users—through gamification strategies [5, 9, 10]. Despite this progress, child-focused assessment tools remain limited. Many existing automatic systems are trained on adult speech and show reduced performance when applied to children [3, 7], whereas complex interfaces continue to hinder clinical adoption [5].

These challenges are connected to broader scientific advances in speech-based diagnosis. Automated analysis has evolved from handcrafted acoustic features toward deep models that learn representations directly from the signal. Classical Machine Learning (ML) approaches have shown accuracies above 85% in the detection of neurological diseases [11], whereas recurrent neural networks have modeled sequential patterns with accuracies near 92% [12]. Convolutional networks and pre-trained representations have demonstrated their capacity to extract clinically meaningful acoustic cues in cognitive and motor-speech disorders [13, 14].

Similarly, telerehabilitation systems offering remote therapy, automatic assessment, and personalized feedback have been successfully validated [15], with deep models contributing to dysarthria rehabilitation and mobile linguistic therapies proving feasible for children with developmental disorders [16, 17].

Beyond childhood speech disorders, automated speech analysis has also been explored in other clinical domains, including neurological conditions, where deep learning models have demonstrated strong performance in extracting clinically relevant acoustic patterns [18-24]. While these studies illustrate the broader potential of speech-based diagnostic technologies, their target populations, speech characteristics, and clinical aims differ from those of pediatric phonetic–phonological assessment. This distinction underscores the need for child-specific datasets, models, and evaluation strategies, which remain comparatively underexplored in the current literature.

Recent advances in self-supervised speech representations, particularly Wav2Vec 2.0, have opened new opportunities for clinical speech analysis by enabling robust acoustic modeling with limited labeled data. Emerging work has demonstrated the applicability of these models to children's speech, including automated phonetic transcription and forced alignment for the assessment of speech sound disorders [25, 26]. These approaches highlight the potential of deep acoustic representations to support objective and scalable assessment in pediatric speech pathology, while also revealing the need for task-specific adaptation and integration with clinically meaningful linguistic information.

Automatic speech-assessment methods have also expanded to multilingual populations, mild cognitive-impairment detection, and therapy recommendation systems [27-29]. Diagnostic tools based on telephone recordings have shown feasibility for Parkinson's disease screening [30]. Within the specific context of children's speech, end-to-end acoustic-linguistic models have reached F-measure values up to 66% in pronunciation-error detection tasks [31], whereas speech reconstruction techniques have improved detection of errors in unseen sentences, strengthening generalization [32]. Additional work has shown that combining prosodic and spectral acoustic attributes (Pitch, Formants, MFCC, LPCC) with classifiers such as J48, MLP, Rotation Forest, and Logit Boost can reach up to 95.2% accuracy even under Autism Spectrum Disorder (ASD)-related noise [33], underscoring the importance of robust and noise-resilient feature representations. Finally, previous studies by the authors compared deep-learning architectures (CNN, RNN, LSTM) and classical models (SVM, Random Forest) for phonological-error classification in children, identifying a Deep Neural Network (DNN) as the most effective, surpassing 94% accuracy in training and establishing the empirical foundation for the final model implemented in this work [34].

To address the limited availability of child-centered, language-specific, and clinically deployable tools for phonetic–phonological assessment, this paper proposes a gamified web-based platform for the automatic diagnosis of phonetic–phonological disorders in children using deep learning. The system integrates deep neural models capable of detecting articulatory errors directly from audio samples with an interactive game designed to sustain children's engagement, while enabling remote operation through the browser without specialized hardware.

II. SYSTEM ARCHITECTURE

The overall architecture of the platform is presented in Figure 1 and consists of five modular layers that interact to support both the children's gamified evaluation process and the therapist's management workflow.

A. Presentation Layer

This layer includes two independent React-based web applications deployed on Netlify.

1) Therapist Portal

The Therapist Portal is the main clinical interface of the platform and supports the configuration, supervision, and validation of automated speech evaluations. The portal is designed to ensure that all assessment processes remain under clinician control, positioning the system as a decision-support tool that assists therapists in the diagnostic process rather than an autonomous diagnostic solution.

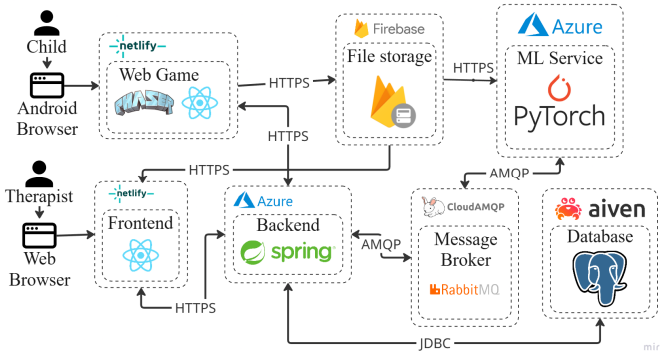

Fig. 1. Overall architecture of the proposed gamified web platform, showing the interaction between presentation, backend services, messaging, ML, and persistence layers.

The workflow begins with patient registration, followed by the creation of a new evaluation session. During configuration, the therapist selects the target phonemes to be assessed using structured multi-selection controls organized by phonetic segments. Default phoneme selections are automatically suggested based on the child's age, whereas full manual adjustment is allowed to accommodate individual clinical needs. An example of the evaluation configuration interface is shown in Figure 2.

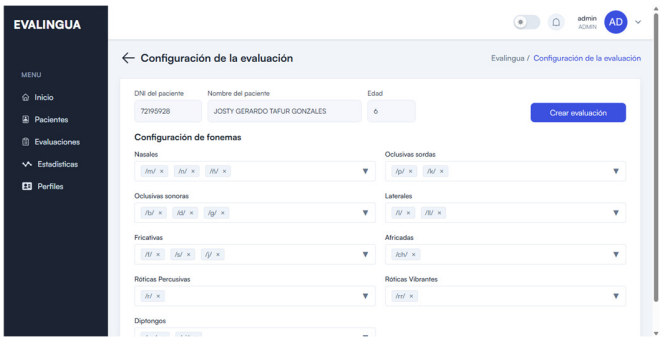

Fig. 2. Therapist Portal interface for evaluation configuration, showing phoneme selection organized by phonetic segments with age-based default suggestions.

Once the evaluation is started, the system generates a QR code that transfers the configured parameters to the gamified web application. After the game-based session is completed, evaluation results are progressively returned to the portal as they are processed by the deep learning module. The therapist can review results through a dedicated interface that presents phoneme-level classifications alongside the corresponding audio recordings, enabling direct validation of the model's outputs.

Based on the aggregated results, the platform generates a structured diagnostic summary organized by phonetic segment, showing phonetic alteration, phonological alteration, or typical production patterns. In addition, the portal offers a longitudinal patient view that allows therapists to check progress across sessions using graphical summaries. An example of the patient results interface is illustrated in Figure 3.

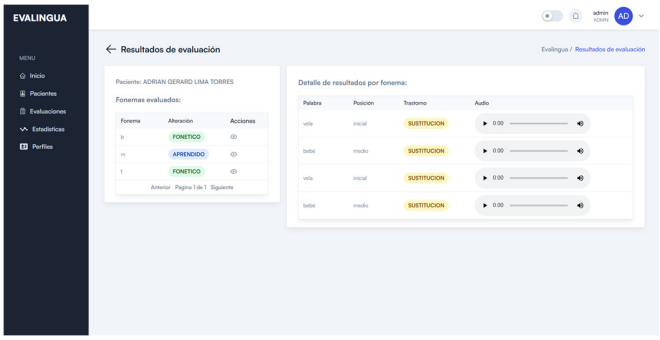

Fig. 3. Patient results view in the Therapist Portal, displaying phoneme-level classification outcomes with corresponding evaluated words and audio recordings.

2) *Web Game*

The Web Game serves as the child-facing part of the platform and functions as an interactive interface for speech sample elicitation. It is accessed via a standard web browser and is dynamically configured for each evaluation session using the QR code generated by the Therapist Portal.

The game employs a zoo-themed visual metaphor in which animal characters guide the child through short interactive challenges designed to elicit target words selected by the therapist. Each challenge requires the child to interact with on-screen objects and pronounce their names, thereby producing speech samples for phonetic–phonological analysis. The game design emphasizes simplicity, positive reinforcement, and non-competitive mechanics to sustain attention without introducing cognitive or emotional pressure.

For each target word, the child is allowed up to three pronunciation attempts. Regardless of pronunciation accuracy, the task progresses to ensure that the child is not penalized for speech difficulties. Immediate visual and auditory feedback is provided after each attempt to keep engagement and motivation within a calm therapeutic environment.

All speech recordings are captured directly through the browser using the Azure Speech API and stored in Firebase Storage. Evaluation metadata are transmitted to the backend in real time, enabling asynchronous processing by the deep learning module. Classification results are returned to the Therapist Portal progressively as analysis is completed, closing the evaluation loop between child interaction and clinician review. An example of the game interface, illustrating the zoo-themed visual metaphor, interactive objects, and bubble-based pronunciation tasks, is shown in Figure 4.

B. *Business Layer*

The backend, implemented in Spring Boot and hosted on Azure Web App, orchestrates system logic, manages authentication and authorization using JSON Web Tokens (JWT), exposes secure REST endpoints for interaction with the web applications, and handles all operational workflows. Each new audio result triggers the generation of an Advanced Message Queuing Protocol (AMQP) message containing its metadata, which is published to the "evaluations" queue in RabbitMQ.

Fig. 4. Gamified web-based evaluation interface, illustrating the zoo-themed design used to elicit target word pronunciations during pediatric speech assessment.

C. Asynchronous Messaging Layer

To decouple interaction from processing, the platform uses RabbitMQ (CloudAMQP). The backend publishes messages to the "evaluations" queue; the ML service consumes them, processes the audio, and publishes classification outputs to the "results" queue. The backend then updates the database accordingly.

D. Machine Learning Layer

Running as an independent Python microservice on Azure Web App, this layer integrates trained models, linguistic encoders, and a Wav2Vec2-based feature extractor. Its internal pipeline includes: receiving evaluation messages, downloading audio from Firebase Storage, preprocessing and feature extraction, classification of phonetic–phonological errors, and publishing results to the backend through the messaging layer.

E. Persistence Layer

A managed PostgreSQL instance (Aiven) stores user profiles, metadata of each recording, and evaluation results. The backend interacts with the database through JPA/Hibernate, ensuring structured persistence and domain integrity.

III. METHODOLOGY

A. Dataset

The dataset used in this study was specifically created for the automatic assessment of phonetic–phonological disorders in preschool children and includes voice recordings collected in real clinical and educational contexts [35]. Data acquisition was conducted in two stages. In the first stage, a clinical corpus of 175 recordings was collected from 7 children aged between 4 and 5 years. In the second stage, three additional participants were incorporated to increase phonetic variability and support model validation, resulting in a final dataset comprising a total of 10 children. All experimental results reported in this study are based on this final dataset.

Recordings were obtained using Android-based browsers under controlled acoustic conditions typically found in therapeutic and educational environments. All samples were stored as uncompressed WAV files with a sampling rate of 16 kHz and a resolution of 16 bits.

To mitigate class imbalance and increase robustness, data augmentation techniques were applied during training, including pitch shifting, speed modification, artificial reverberation, and additive noise. Dataset partitioning followed a stratified strategy balancing phoneme, phonetic position, linguistic segment, and error type. The final split consisted of 70% training, 15% validation, and 15% test data.

Following augmentation, the dataset included approximately 700 audio samples derived from the original recordings. Augmentation was applied exclusively to the training subset to avoid data leakage and preserve the integrity of the evaluation. Although the number of unique participants remains limited, this strategy provided sufficient phoneme-level variability to support a feasibility-oriented evaluation under realistic clinical data constraints.

It is important to note that data partitioning was performed at the sample level with stratification by phoneme and error type. Due to the limited cohort size, a fully speaker-independent split could not be enforced. Consequently, recordings from two children appear in both training and test subsets, although originating from distinct recording sessions and without duplication or augmented variants shared across subsets. This strategy reflects common constraints in early-stage pediatric speech research and was adopted to maximize data use while keeping evaluation integrity. A descriptive summary of the final dataset is presented in Table I.

TABLE I. DESCRIPTIVE SUMMARY OF THE FINAL DATASET USED IN THE EXPERIMENTS

Characteristics	Values
Total number of recordings collected	175
Total number of recordings after augmentation	700
Number of participants (children)	10
Age range	4–5 years old
Sampling frequency	16 kHz
File format	WAV, 16 bits
Collection methods	Direct recording / manual upload
Phonological positions covered	Initial, medial, final
Types of labeled errors	Omission, substitution, distortion, correctness
Applied augmentation technique	Noise, pitch, speed, reverberation
Partition criterion	Balance by type of error/phoneme/segment/position
Data division	70% training, 15% validation, 15% test

B. Model

The proposed speech analysis model combines deep acoustic representations extracted using a pre-trained Wav2Vec2 network with learned linguistic embeddings encoding phoneme identity, phoneme position, and linguistic segment information. This hybrid acoustic–linguistic design enables the system to capture fine-grained phonetic characteristics while incorporating structured symbolic context relevant to speech sound disorder classification. An overview of the processing pipeline is illustrated in Figure 5.

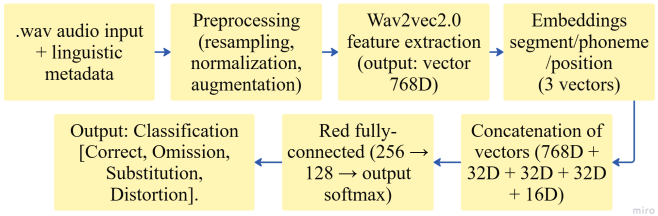

Fig. 5. Overview of the proposed speech-processing model integrating Wav2Vec2-based acoustic features with linguistic embeddings and a deep neural classifier.

1) Signal Preprocessing

All audio samples were standardized to mono signals with a sampling rate of 16 kHz. During training, a probabilistic augmentation module was applied with a 70% probability to further improve robustness and generalization, while avoiding leakage across data subsets.

2) Feature Extraction

Acoustic representations were extracted using the Wav2Vec2FeatureExtractor from the "facebook/wav2vec2-base-960h" model, generating 768-dimensional embeddings for each audio sample. In parallel, symbolic linguistic information was encoded using learnable embeddings, including 32-dimensional phoneme embeddings, 32-dimensional segment embeddings, and 16-dimensional position embeddings. These representations were derived from label-encoded metadata associated with each recording.

3) Classifier Fusion and Architecture

The concatenated acoustic and linguistic vectors were fed into a deep dense neural classifier consisting of two hidden layers with 256 and 128 units, respectively, using ReLU activation and dropout regularization (rate = 0.3). A final softmax layer produced probability estimates for the four target categories: omission, substitution, distortion, and correct production. Most layers of the Wav2Vec2 encoder were kept frozen during training to reduce overfitting. The model was implemented in PyTorch and optimized using the AdamW algorithm.

In addition to the proposed deep learning architecture, several baseline and comparative classifiers were evaluated, including CNN, RNN, LSTM, SVM, Random Forest, and CART models. The architectural configurations, hyperparameter settings, and training procedures for these comparative models were not redefined in the present study. Instead, they were implemented following exactly the same topologies and experimental settings reported in our previous work [34], where a comprehensive comparative analysis of these architectures for childhood phonetic–phonological disorder classification was conducted.

In the current manuscript, these models are re-evaluated exclusively to assess their performance under an expanded and more diverse dataset, allowing a fair comparison focused on generalization behavior rather than architectural optimization. This approach ensures methodological consistency while avoiding redundancy in the presentation of previously published technical details.

C. Training

A two-stage training strategy was employed to evaluate robustness and improve final performance. In the first stage, a 5-fold stratified cross-validation procedure was conducted, ensuring balanced representation of phoneme types and error categories across folds. Training used a class-weighted CrossEntropyLoss to compensate for residual class imbalance, with a learning rate of 5×10^{-5}, linear warm-up and decay scheduling, gradient clipping with a norm threshold of 1.0, and a batch size of 4. The best-performing model from each fold was kept.

In the second stage, the model was retrained using 85% of the available data (training and validation combined), reserving the remaining 15% for final validation. Training was conducted over 50 epochs, after which the final model was exported to the ONNX format to support efficient deployment within the web-based platform.

D. Evaluation Metrics

Accuracy was selected as the primary evaluation metric due to its interpretability and relevance in clinical decision-support scenarios. In routine speech therapy practice, clinicians primarily need a reliable global sign of whether an automated assessment aligns with expert judgment, rather than distribution-sensitive metrics that may be unstable under severe class imbalance.

Although class imbalance was present—particularly affecting distortion-related errors—accuracy was intentionally prioritized at this feasibility stage. Metrics such as macro-averaged F1-score or balanced accuracy may be disproportionately influenced by small class counts and could therefore lead to misleading interpretations in early-stage evaluations. For these reasons, accuracy was considered the most appropriate metric for assessing system reliability in the current exploratory validation context. For completeness, class-wise performance is later analyzed in the Results section to contextualize accuracy outcomes and find categories requiring further data expansion.

IV. RESULTS

A. Deep Learning Model Performance

The performance of the proposed deep learning model was evaluated using the final pediatric speech dataset, including approximately 700 augmented audio samples derived from recordings of 10 children. Figure 6 presents a comparison of overall accuracy across the evaluated classifiers, including classical ML approaches and deep neural architectures. The deep dense neural network achieved the highest accuracy of 83.57%, outperforming both classical and recurrent models.

Computational efficiency was additionally assessed by measuring training time under identical experimental conditions. As shown in Figure 7, classical models showed lower training costs, whereas deep learning architectures required higher computational resources. Nevertheless, the superior classification performance achieved by the deep dense neural model justifies this trade-off in clinical decision-support contexts.

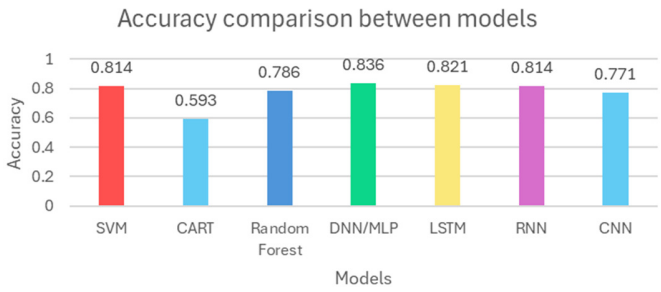

Fig. 6. Accuracy comparison among evaluated classification models using the final pediatric speech dataset.

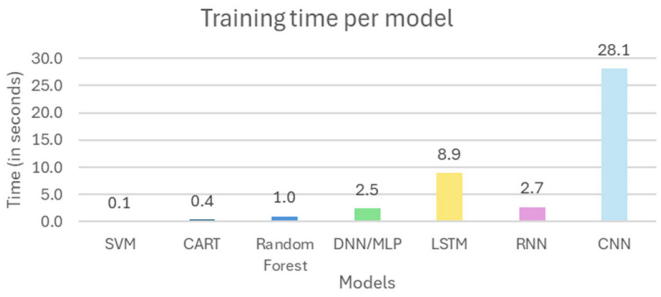

Fig. 7. Training time required by each evaluated model during cross-validation.

Training stability and convergence behavior were evaluated by monitoring loss and accuracy curves throughout training. Figure 8 shows that the deep dense neural network exhibits smoother convergence and lower oscillation compared to other deep learning architectures, showing improved optimization stability.

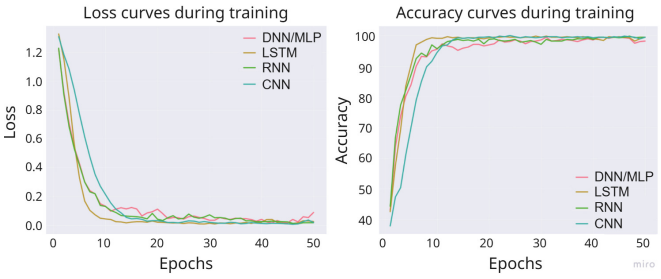

Fig. 8. Training loss and accuracy curves of the evaluated deep learning models.

Finally, Table II compares the results obtained in this study with previous work by the authors [34]. While overall accuracy values decreased due to the increased dataset size and variability, the comparison highlights improved generalization and more realistic performance estimation under clinically representative conditions.

B. Technical Performance

Technical performance was evaluated through controlled load-generation experiments simulating concurrent therapist workflows, including audio upload, result retrieval, and dashboard navigation. These tests allowed reproducible measurements of system latency and throughput.

The average processing latency per audio sample during evaluations was 872 ms, remaining below the predefined upper limit of 900 ms and ensuring smooth system interaction. In addition, the platform demonstrated stable operation under up to 100 simultaneous user sessions, exceeding the minimum success criterion of 50 concurrent sessions. These results confirm the robustness of the proposed architecture and its suitability for real-world clinical deployment.

TABLE II. ACCURACY COMPARISON BETWEEN CURRENT STUDY AND PREVIOUS WORK

Model/algorithm	Accuracy	
	Previous work	Current study
DNN	95.35%	83.57%
CNN	92.67%	77.14%
LSTM	51.10%	82.14%
RNN	45.01%	81.43%
SVM	72.5%	81.43%
Random Forest	70.9%	78.57%
CART	59.67%	59.29%

C. Diagnostic Accuracy of the Model

The diagnostic accuracy of the proposed system was evaluated across 160 clinical scenarios distributed among different phonetic–phonological disorder categories. The overall accuracy achieved was 83.57%, exceeding the predefined minimum clinical threshold of 80%, which is commonly reported in the literature as acceptable for decision-support tools in speech and language assessment rather than for autonomous diagnosis [4, 5]. Figure 9 presents the classification accuracy obtained for each disorder category. Accuracy reached 82.61% for omission, 84.21% for substitution, and 66.67% for distortion cases. The lower performance observed for distortion errors is primarily attributable to the limited availability of representative samples, which restricts the model's ability to learn robust distortion-specific acoustic patterns. Accuracy for correct productions reached 94.62%, although this metric is secondary from a diagnostic standpoint.

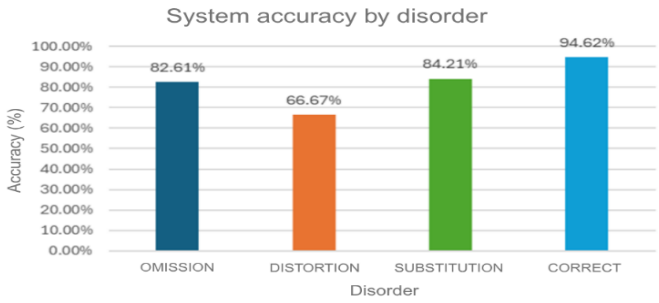

Fig. 9. Classification accuracy by phonetic–phonological category on the final evaluation set.

D. Effectiveness in Real-World Environment

To assess practical applicability, the proposed system was deployed in real clinical settings and compared against traditional manual assessment procedures. Figure 10 illustrates the total evaluation time required by each approach under realistic usage conditions.

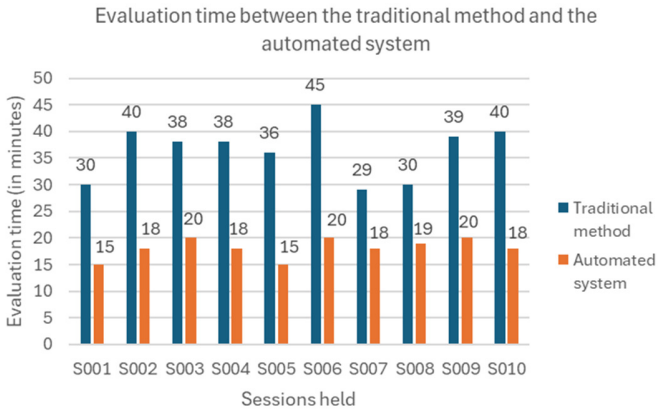

Fig. 10. Comparison of total assessment time between traditional clinical evaluation and the proposed gamified web-based system.

The gamified web-based system achieved a 49.6% reduction in total assessment time, significantly exceeding the minimum improvement threshold of 20%. This reduction proves the system's capacity to streamline therapeutic workflows and reduce clinician workload while supporting diagnostic reliability.

E. Perception of Usability

Usability was assessed using the System Usability Scale (SUS) questionnaire, completed by four certified speech therapists across two clinical sites. The system obtained an average SUS score of 78.5, corresponding to a "good usability" rating according to established international benchmarks.

In addition, functional evaluations based on predefined user stories yielded average ratings between 4.0 and 5.0 out of 5.0, reflecting positive feedback about interface clarity, ease of configuration, and engagement during gamified evaluation sessions. These results show strong acceptance of the platform among clinicians and support its practical integration into therapeutic workflows.

V. DISCUSSION

The findings of this study demonstrate the technical feasibility and clinical relevance of the proposed system for the automated diagnosis of childhood phonetic–phonological disorders. The achieved overall accuracy of 83.57% confirms the suitability of the model as a decision-support tool for speech therapists, exceeding the commonly accepted clinical threshold of 80% reported in related literature [4, 5]. These results align with prior evidence supporting the effectiveness of deep learning approaches for speech impairment analysis, while extending their applicability to a therapist-centered, gamified evaluation environment.

A key contribution of this work lies in the evaluation using an expanded dataset of more than 700 samples, compared to approximately 404 augmented recordings used in the authors' previous study [34]. Although the increased dataset variability led to a moderate reduction in accuracy for some architectures (e.g., DNN from 95.35% to 83.57%), this behavior is consistent with observations reported in [7] and [8], where performance decreases are expected when models are assessed under more realistic and heterogeneous conditions. Consequently, the results presented here provide a more robust and representative estimate of real-world performance.

From a system perspective, the technical evaluation confirms that the proposed platform meets the operational requirements for real-time clinical use. The average processing latency of 872 ms per audio sample and stable operation under up to 100 concurrent user sessions demonstrate that the architecture can support multi-patient environments such as schools and clinical centers, in line with performance recommendations for digital speech-therapy tools [4].

Usability results further support the practical applicability of the system. The mean SUS score of 78.5 indicates good usability and reflects a positive perception among therapists, consistent with previous studies highlighting the benefits of integrating playful dynamics into digital therapeutic tools [9]. Therapists particularly valued the clarity of the outputs and the increased engagement observed during game-based evaluations, suggesting that the interface effectively supports both clinical decision-making and child motivation.

From a clinical safety standpoint, the platform is explicitly designed as a decision-support system rather than an autonomous diagnostic tool. All automated classifications are reviewed and validated by therapists before being incorporated into patient records, and the system allows manual correction of model outputs. In addition, backend audit logs ensure traceability of evaluations and modifications, reducing the risk of misinterpretation of automated results.

Despite these promising outcomes, several limitations must be acknowledged. The lower accuracy observed for distortion errors (66.67%) is primarily attributable to the limited availability of representative samples for this category, restricting the model's ability to learn robust distortion-specific acoustic patterns. This class imbalance also constrains the reliability of distribution-sensitive metrics, supporting the decision to prioritize overall accuracy at this exploratory stage. Furthermore, the partial overlap of speakers between training and evaluation subsets may introduce optimistic bias, a common limitation in early-stage pediatric speech datasets.

Future work will therefore focus on expanding the dataset with a larger and more diverse pediatric population to enable fully speaker-independent validation. This will allow a more rigorous assessment of generalization and support the inclusion of additional evaluation metrics such as macro-averaged F1-score and class-wise recall. Further efforts will also address usability refinements for smaller-screen devices and improve model performance for underrepresented error categories. Overall, the results establish a solid foundation for the continued development and clinical integration of the proposed platform.

VI. CONCLUSION

The present study demonstrated the technical feasibility and practical potential of integrating deep learning models into a gamified web-based platform for the assisted diagnosis of phonetic–phonological disorders in children. The system achieved accuracy levels above the predefined clinical

threshold, reduced evaluation time compared with traditional manual procedures, and was positively evaluated in terms of usability by therapists, confirming its suitability for therapeutic contexts.

Although the results are promising, certain limitations remain. The dataset does not yet provide full phoneme coverage, and the lower accuracy observed in the detection of distortion errors reflects the scarcity of representative samples for this category. Additionally, minor usability issues were found on devices with reduced screen resolution, suggesting opportunities for interface refinement.

Future work will focus on expanding the linguistic and clinical diversity of the dataset, improving model performance for underrepresented disorder types, and integrating the platform with broader clinical management systems. These efforts aim to strengthen the system's applicability and ease its adoption in large-scale therapeutic environments.

ACKNOWLEDGMENT

The authors acknowledge the Dirección de Investigación of the Universidad Peruana de Ciencias Aplicadas (UPC) for supporting this work through the UPC-EXPOST-2026-1 incentive.

REFERENCES

[1] A. K. Namasivayam, D. Coleman, A. O'Dwyer, and P. van Lieshout, "Speech Sound Disorders in Children: An Articulatory Phonology Perspective," *Frontiers in Psychology*, vol. 10, Jan. 2020, Art. no. 2998, https://doi.org/10.3389/fpsyg.2019.02998.

[2] L. Sices, H. G. Taylor, L. Freebairn, A. Hansen, and B. Lewis, "Relationship Between Speech-Sound Disorders and Early Literacy Skills in Preschool-Age Children: Impact of Comorbid Language Impairment," *Journal of Developmental & Behavioral Pediatrics*, vol. 28, no. 6, pp. 438–447, Dec. 2007, https://doi.org/10.1097/DBP.0b013e31811ff8ca.

[3] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer Speech & Language*, vol. 50, pp. 62–84, July 2018, https://doi.org/10.1016/j.csl.2017.12.006.

[4] Z. Brahmi, M. Mahyoob, M. Al-Sarem, J. Algaraady, K. Bousselmi, and A. Alblwi, "Exploring the Role of Machine Learning in Diagnosing and Treating Speech Disorders: A Systematic Literature Review," *Psychology Research and Behavior Management*, vol. 17, pp. 2205–2232, Dec. 2024, https://doi.org/10.2147/PRBM.S460283.

[5] G. A. Attwell, K. E. Bennin, and B. Tekinerdogan, "A Systematic Review of Online Speech Therapy Systems for Intervention in Childhood Speech Communication Disorders," *Sensors*, vol. 22, no. 24, Dec. 2022, Art. no. 9713, https://doi.org/10.3390/s22249713.

[6] Y.-M. Kuo, S.-J. Ruan, Y.-C. Chen, and Y.-W. Tu, "Deep-Learning-Based Automated Classification of Chinese Speech Sound Disorders," *Children*, vol. 9, no. 7, July 2022, Art. no. 996, https://doi.org/10.3390/children9070996.

[7] S. S. Sung, J. So, T.-J. Yoon, and S. Ha, "Automatic detection of speech sound disorder in children using automatic speech recognition and audio classification," *Phonetics and Speech Sciences*, vol. 16, no. 3, pp. 87–94, 2024, https://doi.org/10.13064/KSSS.2024.16.3.087.

[8] X. Zhang, F. Qin, Z. Chen, L. Gao, G. Qiu, and S. Lu, "Fast screening for children's developmental language disorders via comprehensive speech ability evaluation—using a novel deep learning framework," *Annals of Translational Medicine*, vol. 8, no. 11, pp. 707–707, June 2020, https://doi.org/10.21037/atm-19-3097.

[9] T. Brackenbury and L. Kopf, "Serious Games and Gamification: Game-Based Learning in Communication Sciences and Disorders," *Perspectives of the ASHA Special Interest Groups*, vol. 7, no. 2, pp. 482–498, Apr. 2022, https://doi.org/10.1044/2021_PERSP-21-00284.

[10] A. Vaezipour, J. Campbell, D. Theodoros, and T. Russell, "Mobile Apps for Speech-Language Therapy in Adults With Communication Disorders: Review of Content and Quality," *JMIR mHealth and uHealth*, vol. 8, no. 10, Oct. 2020, Art. no. e18858, https://doi.org/10.2196/18858.

[11] A. Iyer *et al.*, "A machine learning method to process voice samples for identification of Parkinson's disease," *Scientific Reports*, vol. 13, no. 1, Nov. 2023, Art. no. 20615, https://doi.org/10.1038/s41598-023-47568-w.

[12] S. Cho *et al.*, "Automatic classification of AD pathology in FTD phenotypes using natural speech," *Alzheimer's & Dementia*, vol. 20, no. 5, pp. 3416–3428, May 2024, https://doi.org/10.1002/alz.13748.

[13] F. García-Gutiérrez *et al.*, "Unveiling the sound of the cognitive status: Machine Learning-based speech analysis in the Alzheimer's disease spectrum," *Alzheimer's Research & Therapy*, vol. 16, no. 1, Feb. 2024, Art. no. 26, https://doi.org/10.1186/s13195-024-01394-y.

[14] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Communication*, vol. 158, Mar. 2024, Art. no. 103047, https://doi.org/10.1016/j.specom.2024.103047.

[15] G. Vuong, C. L. Burns, J. Dignam, D. A. Copland, H. Wedley, and A. J. Hill, "Configuration of a telerehabilitation system to deliver a comprehensive aphasia therapy program via telerehabilitation (TeleCHAT): A human-centred design approach," *Aphasiology*, vol. 39, no. 1, pp. 93–124, Jan. 2025, https://doi.org/10.1080/02687038.2024.2314328.

[16] D. Mulfari, D. La Placa, C. Rovito, A. Celesti, and M. Villari, "Deep learning applications in telerehabilitation speech therapy scenarios," *Computers in Biology and Medicine*, vol. 148, Sept. 2022, Art. no. 105864, https://doi.org/10.1016/j.compbiomed.2022.105864.

[17] A. E. O. Castellanos, C.-M. Liu, and C. Shi, "Deep Mobile Linguistic Therapy for Patients with ASD," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, Oct. 2022, Art. no. 12857, https://doi.org/10.3390/ijerph191912857.

[18] A. S. Nunes *et al.*, "Digital assessment of speech in Huntington disease," *Frontiers in Neurology*, vol. 15, Jan. 2024, Art. no. 1310548, https://doi.org/10.3389/fneur.2024.1310548.

[19] Y. Momota *et al.*, "Language patterns in Japanese patients with Alzheimer disease: A machine learning approach," *Psychiatry and Clinical Neurosciences*, vol. 77, no. 5, pp. 273–281, May 2023, https://doi.org/10.1111/pcn.13526.

[20] G. Gosztolya, V. Svindt, J. Bóna, and I. Hoffmann, "Extracting Phonetic Posterior-Based Features for Detecting Multiple Sclerosis From Speech," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3234–3244, 2023, https://doi.org/10.1109/TNSRE.2023.3300532.

[21] B. G. Schultz *et al.*, "Disease Delineation for Multiple Sclerosis, Friedreich Ataxia, and Healthy Controls Using Supervised Machine Learning on Speech Acoustics," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4278–4285, 2023, https://doi.org/10.1109/TNSRE.2023.3321874.

[22] J. Liu *et al.*, "Efficient Pause Extraction and Encode Strategy for Alzheimer's Disease Detection Using Only Acoustic Features from Spontaneous Speech," *Brain Sciences*, vol. 13, no. 3, Mar. 2023, Art. no. 477, https://doi.org/10.3390/brainsci13030477.

[23] M. Geng *et al.*, "Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2597–2611, 2022, https://doi.org/10.1109/TASLP.2022.3195113.

[24] J. Song *et al.*, "Detection and differentiation of ataxic and hypokinetic dysarthria in cerebellar ataxia and parkinsonian disorders via wave splitting and integrating neural networks," *Plos One*, vol. 17, no. 6, June 2022, Art. no. e0268337, https://doi.org/10.1371/journal.pone.0268337.

[25] P. Kadambi, T. J. Mahr, K. C. Hustad, and V. Berisha, "A Tunable Forced Alignment System Based on Deep Learning: Applications to Child Speech," *Journal of Speech, Language, and Hearing Research*,

vol. 68, no. 7S, pp. 3583–3601, July 2025, https://doi.org/10.1044/2024_JSLHR-24-00347.

[26] Y. Li, D.-S. Pham, R. Ward, N. Hennessey, and T. Tan, "Using AI to Automate Phonetic Transcription and Perform Forced Alignment for Clinical Application in the Assessment of Speech Sound Disorders," in *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*, Philadelphia, PA, USA, 2025.

[27] H. M. D. P. M. Herath, W. A. S. A. Weraniyagoda, R. T. M. Rajapaksha, P. A. D. S. N. Wijesekara, K. L. K. Sudheera, and P. H. J. Chong, "Automatic Assessment of Aphasic Speech Sensed by Audio Sensors for Classification into Aphasia Severity Levels to Recommend Speech Therapies," *Sensors*, vol. 22, no. 18, Sept. 2022, Art. no. 6966, https://doi.org/10.3390/s22186966.

[28] F. Bertini, D. Allevi, G. Lutero, D. Montesi, and L. Calzà, "Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, Oct. 2021, Art. no. 8, https://doi.org/10.1145/3469089.

[29] M. R. Kumar *et al.*, "Dementia Detection from Speech Using Machine Learning and Deep Learning Architectures," *Sensors*, vol. 22, no. 23, Nov. 2022, Art. no. 9311, https://doi.org/10.3390/s22239311.

[30] C. Laganas *et al.*, "Parkinson's Disease Detection Based on Running Speech Data From Phone Calls," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 5, pp. 1573–1584, May 2022, https://doi.org/10.1109/TBME.2021.3116935.

[31] C.-H. Hsiao, S.-J. Ruan, C.-L. Chen, Y.-W. Tu, Y.-C. Chen, and G. M. Rahmatullah, "A Text-Dependent End-to-End Speech Sound Disorder Detection and Diagnosis in Mandarin-Speaking Children," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, 2024, Art. no. 3001911, https://doi.org/10.1109/TIM.2024.3438853.

[32] A. Das and R. Gutierrez-Osuna, "Improving Mispronunciation Detection Using Speech Reconstruction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4420–4433, 2024, https://doi.org/10.1109/TASLP.2024.3434497.

[33] A. Samad, A. U. Rehman, and S. A. Ali, "Performance Evaluation of Learning Classifiers of Children Emotions using Feature Combinations in the Presence of Noise," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 5088–5092, Dec. 2019, https://doi.org/10.48084/etasr.3193.

[34] J. T. Tafur Gonzales, J. B. Bazalar, S. A. Wong Durand, and A. D. García Núñez, "Deep Learning Based Web System for the Automated Diagnosis of Phonological-Phonemic Disorders in Infants," in *2025 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, Nan, Thailand, 2025, pp. 680–685, https://doi.org/10.1109/ECTIDAMTNCON64748.2025.10961963.

[35] J. G. Tafur Gonzales, "Spanish audios classified according to phonetic-phonological speech disorders." Mendeley Data, July 15, 2025, https://doi.org/10.17632/z7dznk98t8.1.

AUTHORS PROFILE

Josty Gerardo Tafur-Gonzales is a Software Engineer at the Peruvian University of Applied Sciences in Lima, Peru (email: U20201C069@upc.edu.pe, ORCID: https://orcid.org/0009-0007-5153-6496).

Joao Arturo Basauri-Bazalar is a Software Engineer at the Peruvian University of Applied Sciences in Lima, Peru (email: U201716123@upc.edu.pe, ORCID: https://orcid.org/0009-0007-4286-0809).

Sandra Wong-Durand has a Master's degree in Artificial Intelligence and a Master's degree in Business Administration from ESAN with focus on Advanced Project Management, She is a Systems Engineer from UNIFE, with specialization studies in Innovation and Leadership at the Escuela Superior de Administración y Dirección de Empresas (ESADE) - Spain, Additional certifications include Process Improvement Management with CMMI at the Software Engineering Institute, Software Quality at UNIFE, Strategic Project Management at PM Certifica, SOA Architectures at IBM and Oracle (email: pcsiswon@upc.edu.pe, ORCID: https://orcid.org/0000-0002-6154-2124).

Pedro Castaneda is a RENACYT Researcher and holds a PhD in Systems Engineering, a Master's degree in Management and Information Technology Management from UNMSM, and a Master's degree in Business Administration (MBA) from ESAN. He leads e-brokerage projects, software development, and process improvement using agile and traditional methodologies. He has the following certifications: Project Management Professional (PMP), Scrum Certified Developer (CSD), IBM Certified Professional in Rational Unified Process, and ORACLE Certifications (Email: pedro.castaneda@untrm.edu.pe, ORCID: https://orcid.org/0000-0003-1865-1293).

Alejandra Onate-Andino holds a degree in Computer Systems Engineering from Escuela Superior Politécnica de Chimborazo (Ecuador), a Master in Network Interconnectivity from Escuela Superior Politécnica de Chimborazo (Ecuador), and a PhD in Systems Engineering and Computer Science from Universidad Mayor de San Marcos (Peru). Currently, she is the Coordinator of the Software Career at the Escuela Superior Politécnica de Chimborazo (Ecuador) (email: monate@espoch.edu.ec).