frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 28 March 2025
DOI 10.3389/fdgth.2025.1552746

OPEN ACCESS

EDITED BY
Mauro Giacomini,
University of Genoa, Italy

REVIEWED BY
Jiayan Zhou,
Stanford University, United States
Zixuan Zhang,
University of Southern California, United States

*CORRESPONDENCE
Matthew E. Pontell
matthew.e.pontell@VUMC.org

RECEIVED 29 December 2024
ACCEPTED 10 March 2025
PUBLISHED 28 March 2025

CITATION
Shirk MU, Dang C, Cho J, Chen H, Hofstetter L, Bijur J, Lucas C, James A, Guzman R-T, Hiller A, Alter N, Stone A, Powell M and Pontell ME (2025) Leveraging large language models for automated detection of velopharyngeal dysfunction in patients with cleft palate.
Front. Digit. Health 7:1552746.
doi: 10.3389/fdgth.2025.1552746

Leveraging large language models for automated detection of velopharyngeal dysfunction in patients with cleft palate

Myranda Uselton Shirk[1], Catherine Dang[1], Jaewoo Cho[1], Hanlin Chen[1], Lily Hofstetter[1], Jack Bijur[1], Claiborne Lucas[2], Andrew James[3], Ricardo-Torres Guzman[3], Andrea Hiller[3], Noah Alter[3], Amy Stone[4], Maria Powell[4] and Matthew E. Pontell[3,5]*

[1]Data Science Institute, Vanderbilt University, Nashville, TN, United States, [2]Department of General Surgery, Prisma Health Greenville, Greenville, SC, United States, [3]Department of Plastic Surgery, Vanderbilt University Medical Center, Nashville, TN, United States, [4]Department of Otolaryngology, Vanderbilt University Medical Center, Nashville, TN, United States, [5]Division of Pediatric Plastic Surgery, Monroe Carell Jr. Children's Hospital, Nashville, TN, United States

Background: Hypernasality, a hallmark of velopharyngeal insufficiency (VPI), is a speech disorder with significant psychosocial and functional implications. Conventional diagnostic methods rely heavily on specialized expertise and equipment, posing challenges in resource-limited settings. This study explores the application of OpenAI's Whisper model for automated hypernasality detection, offering a scalable and efficient alternative to traditional approaches.
Methods: The Whisper model was adapted for binary classification by replacing its sequence-to-sequence decoder with a custom classification head. A dataset of 184 audio recordings, including 96 hypernasal (cases) and 88 non-hypernasal samples (controls), was used for training and evaluation. The Whisper model's performance was compared to traditional machine learning approaches, including support vector machines (SVM) and random forest (RF) classifiers.
Results: The Whisper-based model effectively detected hypernasality in speech, achieving a test accuracy of 97% and an F1-score of 0.97. It significantly outperformed SVM and RF classifiers, which achieved accuracies of 88.1% and 85.7%, respectively. Whisper demonstrated robust performance across diverse recording conditions and required minimal training data, showcasing its scalability and efficiency for hypernasality detection.
Conclusion: This study demonstrates the effectiveness of the Whisper-based model for hypernasality detection. By providing a reliable pretest probability, the Whisper model can serve as a triaging mechanism to prioritize patients for further evaluation, reducing diagnostic delays and optimizing resource allocation.

KEYWORDS

velopharyngeal dysfunction (VPD), hypernasality detection, artificial intelligence (AI), cleft palate, machine learning (ML), speech diagnostics

1 Introduction

Cleft palate affects approximately 1 in 700 live births worldwide and requires surgical intervention during infancy to prevent adverse feeding, speech, and developmental outcomes (1–3). Despite corrective surgery, up to 30% of patients develop velopharyngeal dysfunction (VPD), a speech disorder marked by hypernasality and reduced

intelligibility (4–6). VPD significantly impairs communication and has profound psychosocial consequences (7–9). An accurate diagnosis of VPD relies on a perceptual speech analysis by specialized speech-language pathologists (SLPs), often with adjunctive testing with videonasoendoscopy, nasometry and different types of imaging (10, 11). As such, the diagnosis of VPD is highly dependent on specialized expertise and costly testing equipment. Both factors make VPD care nearly inaccessible in low- and middle-income countries (LMICs). As a result, there is an unknown number of patients who remain undiagnosed and untreated, further perpetuating disparities in care for orofacial cleft patients in LMICs (12–15).

Efforts to increase capacity in the diagnosis and treatment of VPD have harnessed the power of artificial intelligence (AI) and machine learning (ML). These models autonomously conceptualize non-linear relationships in data, making them particularly well-suited for nuanced tasks such as VPD detection. Multiple teams have explored traditional ML approaches using support vector machines (SVMs) and random forest (RF) classifiers, utilizing engineered features like Mel Frequency Cepstral Coefficients (MFCCs) to identify patterns in audio data (16–18). While these methods have demonstrated some effectiveness, their reliance on extensive preprocessing and feature engineering limits their practicality, especially in real-world settings (16–18). Similarly, deep learning models such as convolutional neural networks (CNNs) have shown promise but typically require large, annotated datasets, often amounting to thousands of hours of audio, to achieve clinically meaningful performance (19, 20). Furthermore, many of these models are restricted to analyzing specific phonetic sounds or operate within narrow linguistic contexts, which can hinder their generalizability across heterogenous populations and languages (16–20).

Recent advancements in Large Language Models (LLMs), particularly OpenAI's Whisper model, offer a promising approach to VPD detection by leveraging pre-trained audio processing capabilities (21). Unlike conventional models that require extensive preprocessing and manual feature engineering, Whisper autonomously extracts acoustic data directly from raw audio files, enhancing efficiency and real-world applicability. By utilizing a transformer-based architecture trained on multilingual datasets, Whisper excels at capturing subtle acoustic variations, making it well-suited for detecting hypernasality and other speech irregularities associated with VPD. Its architecture is inherently designed to accommodate diverse linguistic contexts, allowing for seamless integration across varied speech patterns and dialects (21, 22). This versatility is particularly valuable in low- and middle-income countries (LMICs), where linguistic diversity and resource limitations pose significant diagnostic challenges (23). With targeted refinements, Whisper can further enhance existing diagnostic methods, improving accessibility and broadening its clinical utility. Despite this potential, Whisper's utilization in VPD detection remains largely unexplored, presenting an opportunity to advance global healthcare equity through AI-driven speech analysis.

The aim of this study is to leverage Whisper's pre-trained audio processing capabilities to develop a model that can automatically detect the presence of VPD by voice sample alone. We hypothesize that Whisper's key encoded features can be repurposed to identify patterns of VPD within voice samples, with a primary endpoint of model accuracy.

2 Methods

This study was approved by the Institutional Review Board at Vanderbilt University Medical Center/Monroe Carell Jr. Children's Hospital (IRB#212135). Audio samples of patients with a diagnosis of VPD, as well as unaffected voice samples, were sources from publicly available online repositories and institutional datasets to ensure a diverse representation of speech patterns. Unaffected audio samples were sourced from the Centers for Disease Control and Prevention and the Eastern Ontario Health Unit (21, 22). VPD voice samples were obtained from multiple publicly available sources (23–29).

All recordings were preprocessed into WAV format and resampled to 16 kHz to ensure compatibility with the Whisper model. To standardize inputs, each recording was processed to fit Whisper's fixed 30 s input window by zero-padding shorter samples and truncating longer ones. Metadata, including recording conditions and file duration, was cataloged for each sample.

Patient-level variables, including age, sex, and severity of hypernasality, were not included in the analysis due to the lack of this information in the publicly accessible datasets.

2.1 Study design

This study involved data preprocessing, adapting the multilingual Whisper model for binary classification tasks, and comparing its performance against traditional machine learning models. The models evaluated included Whisper-base, Whisper-medium, and Whisper-large-v2, each paired with a custom classification head. Baseline comparisons were conducted using Support Vector Machine (SVM) and Random Forest (RF) classifiers.

2.2 Whisper model variants

All three variants—Whisper-base, Whisper-medium, and Whisper-large-v2—share the same transformer-based architecture but differ in parameter size, which influences their computational efficiency and ability to capture complex speech features. Whisper-base, the smallest model, prioritizes speed but has lower precision. Whisper-medium offers a balance between performance and computational demand, while Whisper-large-v2, the most complex variant, has the highest number of parameters and was trained for additional epochs to improve accuracy (24).

2.3 MFCC extraction for baseline models

For the baseline models, MFCCs were extracted using the LibROSA library in Python (Python Software Foundation,

Wilmington, DE) (25). To ensure consistency across varying recording lengths, the extracted MFCC sequences were mean-aggregated over time, generating a fixed-length feature vector. These processed representations were then used as inputs for the SVM and RF classifiers.

2.4 Model architecture and training

The Whisper model, originally designed for robust speech-to-text transcription, was adapted for binary classification of VPD. This was achieved by replacing its sequence-to-sequence decoder with a custom classification head. (Figure 1) Each encoder processed the audio data, passing the extracted features to a neural network classifier. The classification head is comprised of five fully connected layers with progressively decreasing output dimensions (4096, 2048, 1024, 512, and 2 nodes), employing Rectified Linear Unit (ReLU) activations between layers. (Table 1) A softmax activation function in the final layer produced probabilistic outputs for classification.

To optimize computational efficiency, the pre-trained parameters of the Whisper encoder were frozen, allowing the classification head to focus on learning task-specific features. The model was trained for 10 epochs using the AdamW optimizer, with a learning rate of 0.00002 and weight decay of 0.0005. (Table 2) Cross-entropy loss was used as the objective function, and early stopping with validation monitoring was implemented to prevent overfitting. All training and evaluation were conducted on an NVIDIA DGX A100 GPU.

2.5 Baseline comparisons

Baseline models, including SVM and RF classifiers, were implemented for comparative analysis. These models utilized MFCCs as input features, requiring extensive feature engineering and preprocessing. Identical data splits were used to benchmark the performance of the Whisper-based model against these traditional approaches.

2.6 Performance evaluation

The dataset was randomly divided into training (70%), validation (15%), and test (15%) subsets, maintaining a balanced distribution of VPD and non-VPD samples. Model performance was assessed using metrics such as accuracy, F1-score, and computational efficiency. Validation metrics were monitored during training to identify the best-performing model for final evaluation on the test dataset.

2.7 Software and reproducibility

All experiments were implemented using Python, with PyTorch for model training and the Hugging Face library for accessing Whisper encoders. The codebase, including preprocessing scripts and training pipelines, is available in a publicly accessible GitHub repository to ensure reproducibility.

3 Results

3.1 Dataset characteristics

The dataset included 184 audio samples, with 96 VPD (cases) and 88 non-VPD (controls) recordings. Audio sample durations ranged from 0.44 to 9.35 s. To ensure balanced evaluation across VPD and non-VPD samples, the data was split into training (70%, $n = 129$), validation (15%, $n = 28$), and test (15%, $n = 27$) subsets, maintaining the original 96:88 case-to-control ratio. The final distribution across subsets is shown in Table 3.

3.2 Whisper-Based model performance

The Whisper-based model demonstrated strong performance across configurations. (Table 4) The Whisper-base configuration, paired with a custom classification head, achieved the highest test accuracy of 97.0% and an F1-score of 0.97. Whisper-medium and Whisper-large-v2 configurations achieved test accuracies of 94.9% and 89.2%, with corresponding F1 scores of 0.95 and 0.89.

3.3 Baseline model comparisons

Baseline models trained using MFCCs as input features showed lower performance compared to the Whisper-based models. The SVM model achieved a test accuracy of 88.1% and an F1 score of 0.86, while the RF classifier achieved a test accuracy of 85.7% and an F1 score of 0.88. These traditional models required significant preprocessing and manual feature engineering, which increased computational overhead.

4 Discussion

This study demonstrates the effectiveness of OpenAI's Whisper model for automated VPD detection, achieving a test accuracy of 97% and an F1-score of 0.97. These results significantly outperform baseline models, including SVM (88.1% accuracy) and RF classifiers (85.7% accuracy), which relied on handcrafted features such as MFCCs. Whisper's ability to capture nuanced speech characteristics directly from raw audio samples, coupled with its holistic processing capabilities, underscores its value in both technical performance and clinical utility. These findings validate the feasibility of leveraging ML technology to bridge gaps in diagnostic care, particularly in underserved and resource-constrained settings.

Conventional methods for hypernasality detection rely heavily on perceptual assessments conducted by SLPs and adjunctive tools such as nasometry, videofluoroscopy, or imaging systems (10, 11).

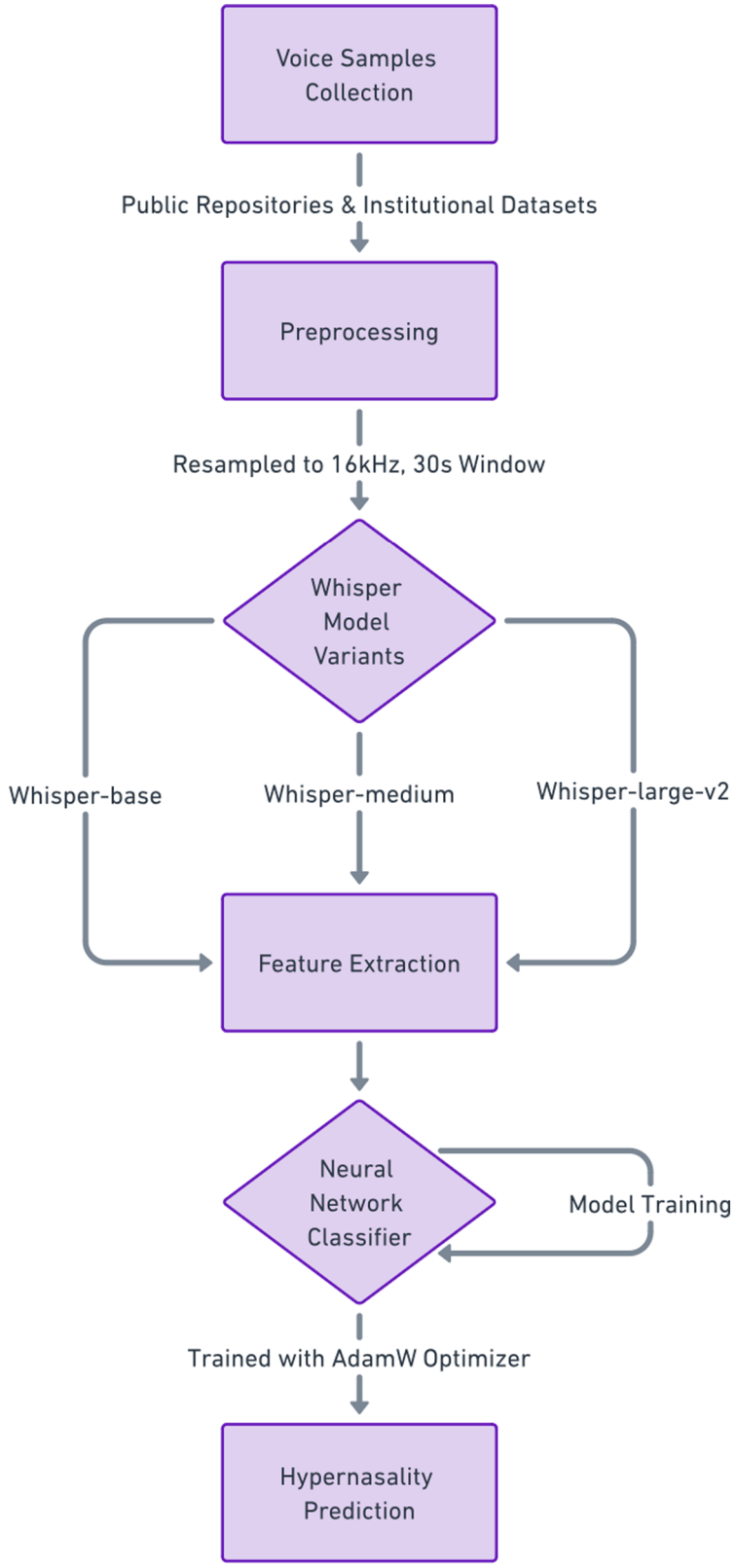

FIGURE 1
Workflow of the whisper-based model for hypernasality detection.

TABLE 1 Neural network architecture for hypernasality detection.

Layer	Operation	Output shape
Linear	Fully connected	4,096
ReLU	Activation	4,096
Linear	Fully connected	2,048
ReLU	Activation	2,048
Linear	Fully connected	1,024
ReLU	Activation	1,024
Linear	Fully connected	512
ReLU	Activation	512
Linear	Fully connected	2
Softmax	Activation	2

TABLE 2 Training optimizer and hyperparameter configuration for model training.

Training optimizer	
Optimizer	Hyperparameters
AdamW	Learning Rate: 0.00002
	β1: 0.95,
	β2: 0.95,
	λ: 0.0005

TABLE 3 Dataset distribution across training, validation, and test sets.

Dataset split	Total samples	VPD samples	Non-VPD samples
Training (70%)	129	67	62
Validation (15%)	28	14	14
Test (15%)	27	15	12
Total	184	96	88

TABLE 4 Comparison of hypernasality detection models.

Model performance		
Model	Test accuracy	F-1 score
Whisper-base + Classifier	97.00%	0.97
Whisper-medium + Classifier	94.90%	0.95
Whisper-large-v2 + Classifier	89.20%	0.89
SVM	88.10%	0.86
RF	85.70%	0.88

While effective, these methods pose substantial barriers due to significant costs, reliance on specialized equipment, and the need for highly trained personnel. These challenges are particularly pronounced in LMICs, where healthcare infrastructure is limited, and access to qualified professionals is often scarce (13, 15, 30, 31). As a result, many patients in these regions remain undiagnosed and untreated, exacerbating the functional and psychosocial burdens associated with VPD (12–15).

The Whisper-based model provides an innovative solution by offering a high pretest probability of VPD, ensuring efficient triage of patients most likely to benefit from specialized care. By reducing unnecessary referrals and diagnostic procedures, the model minimizes financial and operational waste for healthcare providers and families (32, 33). These benefits are particularly relevant in LMICs, where the cost of consultations, procedures, and follow-up care can prohibit access to care (34–36).

The success of the Whisper-based model lies in its technical architecture. Whisper's pre-trained encoder autonomously extracts high-dimensional acoustic features, such as pitch, tone, and resonance, directly from raw audio data, providing a rich foundation for downstream tasks and eliminating the need for extensive preprocessing (21). Unlike traditional models that process segmented audio, Whisper holistically analyzes entire audio samples, enhancing its clinical applicability. By replacing its sequence-to-sequence decoder with a classification head, Whisper's encoded features can be repurposed for binary classification of VPD detection. This modular design not only minimizes computational demands but also preserves the integrity of the learned representations, enabling efficient and accurate classification of VPD. Importantly, the model demonstrated consistent accuracy across diverse recording conditions, underscoring its resilience to variability in data quality, linguistic diversity, and speaker characteristics, a critical attribute for global healthcare applications in LMICs.

Interestingly, Whisper-base outperformed Whisper-large in hypernasality detection, an unexpected finding given the typical advantage of larger models in speech-related tasks. One likely explanation is overfitting, as Whisper-large's greater parameter count may have captured irrelevant speaker variations, background noise, or linguistic structures rather than the core acoustic features of hypernasality. Additionally, because Whisper was originally designed for speech-to-text transcription, larger models may allocate more resources towards linguistic structure and phoneme recognition, which are not directly relevant to hypernasality classification. In contrast, Whisper-base's streamlined architecture may have retained the essential acoustic features necessary for detecting hypernasality without over-prioritizing language modeling. Furthermore, freezing the encoder may have disproportionately affected Whisper-large, as its deeper architecture depends on layer-wise refinements that could have been disrupted. In comparison, Whisper-base may have been inherently better suited for direct acoustic feature extraction, requiring fewer trainable parameters to adapt effectively to the classification task. These findings underscore the importance of model selection and adaptation in AI-driven speech pathology applications.

The mobile integration of the Whisper-based model represents a logical and impactful next step in improving access to VPD detection and care. With the widespread availability of smartphones, deploying this technology on mobile platforms could democratize diagnostic access. A smartphone-based application could record and analyze speech locally, providing immediate feedback to users without requiring an SLP (37). When combined with cloud computing, the model could support large-scale data analysis, enabling personalized diagnostic insights and more comprehensive population health monitoring (38). This approach would facilitate earlier identification of VPD, expediting referrals for surgical or therapeutic interventions. By reducing diagnostic delays, this technology has the potential to improve long-term psychosocial and developmental outcomes for individuals with VPD (39, 40). Additionally, integrating the Whisper-based model into telemedicine platforms could bridge

gaps in care by connecting underserved populations to specialized services (41). This capability empowers community healthcare workers to perform screenings and identify high-risk patients, amplifying the reach of existing healthcare resources.

Despite promising results, this study has several limitations. The dataset was relatively small, consisting of only 184 audio samples, with limited validation and test sets. While the model achieved high accuracy, the small sample size may impact generalizability, particularly across diverse populations, linguistic backgrounds, and recording conditions. Additionally, the study relied exclusively on publicly available data, which may not fully capture the complexity of clinical settings or the variability of patient presentations (42–46). Future research should incorporate proprietary datasets with greater diversity in noise levels, patient demographics, and linguistic contexts. Prospective validation in clinical environments is also needed to assess real-world performance and usability. Additionally, this study did not compare Whisper against a naive neural network, which could provide further insight into the benefits of pre-trained transformer-based models. Exploring this comparison in future research would help contextualize Whisper's performance in hypernasality detection. Lastly, while Whisper's pre-trained encoder demonstrated strong results with English-language samples, additional optimization is necessary to ensure robust performance across non-English languages and dialects, a critical requirement for global scalability.

The strengths of this study lie in its innovative application of OpenAI's Whisper model, which achieves high accuracy and computational efficiency for VPD detection with minimal training data. Additionally, the model's robustness across varying audio conditions makes it highly suitable for real-world deployment. By combining technical innovation with clinical relevance, this study lays the groundwork for deploying intelligent diagnostic tools worldwide, improving care for individuals with cleft-related speech disorders.

5 Conclusion

This study demonstrates the feasibility of adapting OpenAI's Whisper model for automated VPD detection by replacing its sequence-to-sequence decoder with a custom classification head. The adapted model achieved a test accuracy of 97% and an F1-score of 0.97, significantly outperforming traditional models such as support vector machines (accuracy of 88.1%) and random forest classifiers (accuracy of 85.7%). These findings lay the groundwork for future AI-driven tools that can expand access to diagnostic and therapeutic care for cleft-related velopharyngeal dysfunction. AI/ML approaches are particularly suited for care delivery in LMICs, where resources are constrained and clinical expertise is often unavailable.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. CD: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. JC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. HC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. LH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. JB: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. CL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AJ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. R-TG: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. NA: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. MPow: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. MPon: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project

administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors would like to thank the Data Science Institute at Vanderbilt University for providing technical resources and support throughout this project. We also extend our gratitude to the Department of Plastic Surgery and the Division of Pediatric Plastic Surgery at Monroe Carell Jr. Children's Hospital for their guidance and collaboration. Special thanks to the speech-language pathology teams for their expertise and assistance in curating datasets for this study. Additionally, this work benefited from publicly available voice recordings sourced from institutions such as the Centers for Disease Control and Prevention and the Eastern Ontario Health Unit. We acknowledge the role of OpenAI's Whisper framework in facilitating this study's exploration of artificial intelligence applications in speech diagnostics. Finally, we recognize the invaluable contributions of all co-authors and colleagues who reviewed and refined the manuscript. This work has not been previously published but was derived from an ongoing effort to expand access to innovative diagnostics for velopharyngeal dysfunction.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Leslie EJ, Marazita ML. Genetics of cleft lip and cleft palate. *Am J Med Genet C Semin Med Genet*. (2013) 163C(4):246–58. doi: 10.1002/ajmg.c.31381

2. Paradowska-Stolarz A, Mikulewicz M, Duś-Ilnicka I. Current concepts and challenges in the treatment of cleft lip and palate patients-a comprehensive review. *J Pers Med*. (2022) 12(12):2089. doi: 10.3390/jpm12122089

3. Posnick JC, Kinard BE. Challenges in the successful reconstruction of cleft lip and palate: managing the nasomaxillary deformity in adolescence. *Plast Reconstr Surg*. (2020) 145(3):591e–603e. doi: 10.1097/PRS.0000000000006614

4. Smetona JT, Naran S, Ford M, Losee JE. What's new in cleft palate and velopharyngeal dysfunction management: an update. *Plast Reconstr Surg Aug*. (2024) 154(2):378e–90e. doi: 10.1097/PRS.0000000000011312

5. Hopper RA, Tse R, Smartt J, Swanson J, Kinter S. Cleft palate repair and velopharyngeal dysfunction. *Plast Reconstr Surg*. (2014) 133(6):852e–64e. doi: 10.1097/PRS.0000000000000184

6. Sell D, Mildinhall S, Albery L, Wills AK, Sandy JR, Ness AR. The cleft care UK study. Part 4: perceptual speech outcomes. *Orthod Craniofac Res*. (2015) 18(Suppl 2):36–46. doi: 10.1111/ocr.12112

7. Sweeney WM, Lanier ST, Purnell CA, Gosain AK. Genetics of cleft palate and velopharyngeal insufficiency. *J Pediatr Genet*. (2015) 4(1):9–16. doi: 10.1055/s-0035-1554978

8. Barr L, Thibeault SL, Muntz H, de Serres L. Quality of life in children with velopharyngeal insufficiency. *Arch Otolaryngol Head Neck Surg*. (2007) 133(3):224–9. doi: 10.1001/archotol.133.3.224

9. Bhuskute A, Skirko JR, Roth C, Bayoumi A, Durbin-Johnson B, Tollefson TT. Association of velopharyngeal insufficiency with quality of life and patient-reported outcomes after speech surgery. *JAMA Facial Plast Surg*. (2017) 19(5):406–12. doi: 10.1001/jamafacial.2017.0639

10. Brydges HT, Laspro M, Verzella AN, Alcon A, Schechter J, Cassidy MF, et al. Contemporary prevalence of oral clefts in the US: geographic and socioeconomic considerations. *J Clin Med*. (2024) 13(9):2570. doi: 10.3390/jcm13092570

11. Putri FA, Pattamatta M, Anita SES, Maulina T. The global occurrences of cleft lip and palate in pediatric patients and their association with demographic factors: a narrative review. *Children (Basel)*. (2024) 11(3):322. doi: 10.3390/children11030322

12. Xepoleas MD, Naidu P, Nagengast E, Collier Z, Islip D, Khatra J, et al. Systematic review of postoperative velopharyngeal insufficiency: incidence and association with palatoplasty timing and technique. *J Craniofac Surg*. (2023) 34(6):1644–9. doi: 10.1097/SCS.0000000000009555

13. Lucas C, Torres-Guzman R, James AJ, Corlew S, Stone A, Powell ME, et al. Machine learning for automatic detection of velopharyngeal dysfunction: a preliminary report. *J Craniofac Surg*. (2024) 11(6):771. doi: 10.1097/SCS.0000000000010147

14. Peters DH, Garg A, Bloom G, Walker DG, Brieger WR, Rahman MH. Poverty and access to health care in developing countries. *Ann N Y Acad Sci*. (2008) 1136:161–71. doi: 10.1196/annals.1425.011

15. Pantoja T, Opiyo N, Lewin S, Paulsen E, Ciapponi A, Wiysonge CS, et al. Implementation strategies for health systems in low-income countries: an overview of systematic reviews. *Cochrane Database Syst Rev*. (2017) 9(9):CD011086. doi: 10.1002/14651858.CD011086.pub2

16. Rogers HP, Hseu A, Kim J, Silberholz E, Jo S, Dorste A, et al. Voice as a biomarker of pediatric health: a scoping review. *Children (Basel)*. (2024) 11(6):684. doi: 10.3390/children11060684

17. Rong P, Heidrick L, Pattee GL. A multimodal approach to automated hierarchical assessment of bulbar involvement in amyotrophic lateral sclerosis. *Front Neurol*. (2024) 15:1396002. doi: 10.3389/fneur.2024.1396002

18. Dhillon H, Chaudhari PK, Dhingra K, Kuo R, Sokhi RK, Alam MK, et al. Current applications of artificial intelligence in cleft care: a scoping review. *Front Med (Lausanne)*. (2021) 8:676490. doi: 10.3389/fmed.2021.676490

19. Chu HC, Zhang YL, Chiang HC. A CNN sound classification mechanism using data augmentation. *Sensors (Basel)*. (2023) 23(15):6972. doi: 10.3390/s23156972

20. Bhat GS, Shankar N, Panahi IMS. Automated machine learning based speech classification for hearing aid applications and its real-time implementation on smartphone. *Annu Int Conf IEEE Eng Med Biol Soc*. (2020) 2020:956–9. doi: 10.1109/EMBC44109.2020.9175693

21. OpenAI. *Introducing Whisper*. San Francisco, CA: OpenAI (2022). Available at: https://openai.com/index/whisper/ (Accessed March 07, 2025).

22. Radford A, Kim J, Xu T, Brockman G, Mcleavey C, Sutskever I. *Robust Speech Recognition via Large-Scale Weak Supervision*. San Francisco, CA: OpenAI (2022). Available at: https://cdn.openai.com/papers/whisper.pdf (Accessed March 07, 2025).

23. Cowan T, Paroby C, Leibold LJ, Buss E, Rodriguez B, Calandruccio L. Masked-Speech recognition for linguistically diverse populations: a focused review and suggestions for the future. *J Speech Lang Hear Res*. (2022) 65(8):3195–216. doi: 10.1044/2022_JSLHR-22-00011

24. OpenAI. *Whisper-Large v2*. New York, NY: Hugging Face (2023). Available at: https://huggingface.co/openai/whisper-large-v2 (Accessed March 07, 2025).

25. Librosa Development Team. librosa.feature.mfcc. *Librosa 0.10.2 Documentation*. San Francisco, CA: GitHub (2023). Available at: https://librosa.org/doc/latest/generated/librosa.feature.mfcc.html (Accessed March 07, 2025).

26. Centers for Disease Control and Prevention. *CDC's Developmental Milestones*. Atlanta, GA: CDC (2023). Available at: https://www.cdc.gov/ncbddd/actearly/milestones/index.html (Accessed December 01, 2023).

27. Eastern Ontario Health Unit. Let's talk: tips for nuilding your child's speech and language skills (2017). Available at: https://www.youtube.com/watch?v=K0aHjxzDb7I (Accessed December 01, 2023).

28. Fauquier ENT. What is VPI (Velopharyngeal Insufficiency)? (2018). Available at: https://www.youtube.com/watch?v=WM5fVCdBPHs (Accessed March 18, 2025).

29. Jones & Bartlett Learning. Hypernasality (2018). Available at: https://www.youtube.com/watch?v=KWz5_fpnZYc (Accessed December 01, 2023).

30. Aslam MZ, Trail M, Cassell AK 3rd, Khan AB, Payne S. Establishing a sustainable healthcare environment in low- and middle-income countries. *BJU Int*. (2022) 129(2):134–42. doi: 10.1111/bju.15659

31. Meyers D, Brady J, Grace E, Chaves K, Gray D, Barton B, et al. *2021 National Healthcare Quality and Disparities Report*. Rockville, MD: Agency for Healthcare Research and Quality (US) (2021).

32. Yi N, Baik D, Baek G. The effects of applying artificial intelligence to triage in the emergency department: a systematic review of prospective studies. *J Nurs Scholarsh*. (2024) 57(1):105–18. doi: 10.1111/jnu.13024

33. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, et al. A review of the role of artificial intelligence in healthcare. *J Pers Med*. (2023) 13(6):951. doi: 10.3390/jpm13060951

34. Galbraith AA, Wong ST, Kim SE, Newacheck PW. Out-of-pocket financial burden for low-income families with children: socioeconomic disparities and effects of insurance. *Health Serv Res*. (2005) 40(6 Pt 1):1722–36. doi: 10.1111/j.1475-6773.2005.00421.x

35. Yerramilli P, Chopra M, Rasanathan K. The cost of inaction on health equity and its social determinants. *BMJ Glob Health*. (2024) 9(Suppl 1):e012690. doi: 10.1136/bmjgh-2023-012690

36. de Siqueira Filha NT, Li J, Phillips-Howard PA, Quayyum Z, Kibuchi E, Mithu MIH, et al. The economics of healthcare access: a scoping review on the economic impact of healthcare access for vulnerable urban populations in low- and middle-income countries. *Int J Equity Health*. (2022) 21:191. doi: 10.1186/s12939-022-01804-3

37. Mantena S, Celi LA, Keshavjee S, Beratarrechea A. Improving community health-care screenings with smartphone-based AI technologies. *Lancet Digit Health*. (2021) 3(5):e280–2. doi: 10.1016/S2589-7500(21)00054-6

38. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J*. (2021) 8(2):e188–94. doi: 10.7861/fhj.2021-0095

39. Berisha V, Liss JM. Responsible development of clinical speech AI: bridging the gap between clinical research and technology. *NPJ Digit Med*. (2024) 7(1):208. doi: 10.1038/s41746-024-01199-1

40. Pitkänen VV, Alaluusua SA, Geneid A, Vuola PMB, Leikola J, Saarikko AM. How early can we predict the need for VPI surgery? *Plast Reconstr Surg Glob Open*. (2022) 10(11):e4678. doi: 10.1097/GOX.0000000000004678

41. Butzner M, Cuffee Y. Telehealth interventions and outcomes across rural communities in the United States: narrative review. *J Med Internet Res*. (2021) 23(8):e29575. doi: 10.2196/29575

42. LEADERSproject. Cleft palate speech therapy using books for phrases and sentences (2019). Available at: https://www.youtube.com/watch?v=1nHhqdCnwBI (Accessed December 01, 2023).

43. Shriners Children's Chicago. New App may help kids with cleft palate speak easier (2016). Available at: https://www.youtube.com/watch?v=5fubZitvY-Q (Accessed December 01, 2023).

44. Bothel Pediatric and Hand Therapy. Case study: pediatric speech therapy for cleft palate (2016). Available at: https://www.youtube.com/watch?v=noUGRjClUg4 (Accessed December 01, 2023).

45. LEADERSproject. Cleft palate speech and feeding: addressing speech and language before surgery (2016). Available at: https://www.youtube.com/watch?v=-sEt3i0sHr4 (Accessed December 01, 2023).

46. American Speech-Language-Hearing Association. Evaluation and treatment of resonance disorders and velopharyngeal insufficiency (2018). Available at: https://fb.watch/nBEd3Y93AQ/ (Accessed December 01, 2023).