



Speech-Driven Optotype Recognition for Visual Acuity Assessment Using Deep Learning

Dwiny Meidelfi¹, Hendrick², Fanni Sukma³, Ulya Ilhami Arsyah⁴, Jeprizal⁵
^{1,3,4,5}Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia
²Department of Electrical Engineering, Politeknik Negeri Padang, Padang, Indonesia

*dwinymeidelfi@pnp.ac.id

Abstract

Visual acuity (VA) testing is a fundamental procedure for evaluating visual function. Conventional VA examinations require an examiner to interpret patients' responses, which may limit automation and scalability in repeated or remote testing scenarios. This study develops an embedded speech-based optotype recognition system to support digital visual acuity testing workflow. Mel-Frequency Cepstral Coefficients (MFCCs) were used to represent spoken responses, and a Convolutional Neural Network (CNN) was trained to classify nine Snellen optotype letters. Voice recordings were collected from 20 participants, with data augmentation including noise injection, pitch shifting, time stretching, and random gain adjustment to increase acoustic variability. Under the evaluated experimental protocol, the proposed MFCC-CNN model achieved 97.42% accuracy, 94.89% precision, 94.33% recall, and an F1-score of 94.61%, outperforming the evaluated kNN, Random Forest, and SVM baselines. The model was deployed on an NVIDIA Jetson Orin Nano, occupying approximately 8.3 MB and achieving an inference latency of approximately 30–40 ms per prediction. These results demonstrate the technical feasibility of embedded spoken-optotype recognition for supporting visual acuity testing workflows. However, the present study evaluates speech-recognition performance rather than clinical agreement between automatically generated VA scores and clinician-administered measurements. Formal clinical validation and stricter leakage-safe evaluation are therefore required before clinical applicability can be established.

Keywords: CNN; MFCC; snellen chart; speech recognition; visual acuity

How to Cite: D. Meidelfi, H. Hendrick, F. Sukma, U. I. Arsyah, and J. Jeprizal, "Speech-Driven Optotype Recognition for Visual Acuity Assessment Using Deep Learning", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1447–1454, Oct. 2026, doi: [10.29207/resti.v10i5.7538](https://doi.org/10.29207/resti.v10i5.7538).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7538>

Received: January 30, 2026

Accepted: September 5, 2026

Available Online: October 13, 2026

*This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)
Published by [Ikatan Ahli Informatika Indonesia](https://www.lkatanahliinformatika.org/)*

1. Introduction

Visual acuity (VA) assessment is a critical component in evaluating ocular health and detecting visual impairments at an early stage [1]. Traditionally, VA testing has relied on the Snellen Chart, a widely used tool that measures the clarity of vision at a standard distance [2]. Despite its prevalence, this conventional method is inherently susceptible to human error, subjectivity, and inefficiencies, particularly in large-scale or repetitive screening scenarios [3].

To address these challenges, recent advancements in artificial intelligence (AI), machine learning (ML), and sensor technology have paved the way for more objective, automated, and scalable solutions in healthcare diagnostics [4-6]. This study presents the Intelligent Visual Acuity (IVA) Assessment System, an

innovative approach that integrates voice recognition and deep learning techniques to automate VA assessment.

Eye health plays a crucial role in overall quality of life [7], [8], especially in the digital age where prolonged screen exposure contributes to conditions such as Computer Vision Syndrome (CVS), affecting up to 90% of individuals who spend more than three hours daily using digital devices [9]. Symptoms like eye strain, blurred vision, dry eyes, and headaches highlight the importance of early detection and regular VA monitoring [10-12]. Automated tools such as the IVA system may support routine assessment workflows by reducing reliance on manual interpretation of spoken optotype responses.

Several studies have explored the application of AI in ophthalmic diagnostics, including retinal image analysis, cataract detection, and diabetic retinopathy screening [13], [14]. However, research focusing on voice-based automation of VA testing remains limited. Previous efforts at Politeknik Negeri Padang explored K-Means clustering to classify visual acuity decline due to smartphone usage [15]. Building upon this foundation, our work advances the field by integrating audio classification techniques, commonly applied in respiratory disease detection and speech recognition, into visual acuity evaluation [16], [17].

The IVA system leverages Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and employs a Convolutional Neural Network (CNN) for classification [18-23]. By capturing and interpreting spoken responses from patients, the system supports automated recognition of spoken optotype responses during a Snellen-style visual acuity examination. The trained model achieves high accuracy (>97%), validated through confusion matrix analysis and extensive experimentation.

Moreover, the IVA system aligns with global trends toward digital health transformation by supporting semi-automated visual acuity testing. This approach reduces reliance on manual interpretation of patient responses while maintaining compatibility with standard Snellen-based examination procedures. However, the current study focuses on spoken optotype recognition and does not include direct clinical validation of visual acuity measurements. The system is implemented on an embedded Jetson Orin Nano platform to evaluate real-time operation and portability in the proposed testing workflow.

Recent studies have explored digital visual acuity testing systems and tele-ophthalmology platforms that enable remote vision screening using mobile devices and web-based interfaces. These approaches aim to improve accessibility to eye-care services, particularly in underserved areas. In addition, speech-driven human-computer interfaces have been investigated to support hands-free interaction in healthcare environments. Despite these advances, the integration of speech recognition with automated visual acuity testing remains relatively limited.

To further clarify the positioning of the proposed approach, Table 1 summarizes its main characteristics relative to representative digital visual acuity and related technology-assisted assessment approaches reported in the literature. Compared to existing digital visual acuity systems, which primarily rely on manual interaction or touchscreen-based responses, the proposed approach introduces a speech-driven interaction mechanism for optotype recognition. This enables a more accessible and hands-free assessment workflow while maintaining compatibility with standard Snellen-based evaluation procedures.

The contributions of this study are threefold. First, this study proposes an MFCC-CNN pipeline for spoken optotype recognition within a visual acuity testing workflow. Second, the proposed model demonstrates high classification performance for spoken optotype letters through experimental evaluation. Third, the trained model is deployed on an embedded Jetson platform to support real-time and portable testing scenarios.

Table 1. Comparison with Existing Digital Visual Acuity Approaches

Study	Assessment Approach	Interaction Modality	Speech Recognition	Embedded/ Portable Implementation	Clinical Validation
Wanten et al. [3]	Digital/eHealth visual acuity tool	Digital interface	No	Not reported	Evaluated in cataract patients
Baker et al. [24]	Electronic ETDRS visual acuity protocol	Electronic chart/protocol	No	Not reported	Compared with Snellen measurements
Laidlaw et al. [25]	Compact reduced logMAR chart	Chart-based	No	Not reported	Clinical feasibility evaluated
Proposed IVA System	Snellen-based VA workflow	Voice	Yes	Yes	Not yet performed

In summary, this paper presents an embedded speech-driven approach for supporting visual acuity testing through automated optotype recognition. The following sections describe the methodology, experimental results, embedded implementation, and limitations of the evaluated system.

2. Methods

This study followed five research stages. Figure 1 illustrates the sequential research stages undertaken in the development of the Intelligent Visual Acuity (IVA) system. The process begins with problem identification, where the clinical need for a voice-based visual acuity

testing tool is defined as an alternative to traditional chart-based methods.

Following this, the second stage involves dataset acquisition, which includes recording speech samples of optotype letters from participants under controlled conditions. These audio samples form the raw input for the system.

The third stage, signal preprocessing and MFCC extraction, transforms the raw audio into a structured format suitable for machine learning. This step includes denoising, segmentation of the signal into uniform

durations, and extraction of 20 Mel-Frequency Cepstral Coefficients (MFCCs) to capture key spectral features.

In the fourth stage, CNN model design and training is performed. Here, the extracted MFCC features are used to train a Convolutional Neural Network (CNN) model designed for multi-class classification of the spoken letters.

Finally, the fifth stage entails performance evaluation and system deployment, where the trained model is evaluated using standard metrics such as accuracy and recall. Upon achieving satisfactory performance, the system is deployed to an NVIDIA Jetson Orin Nano platform, enabling real-time spoken optotype recognition within the proposed visual acuity testing workflow.

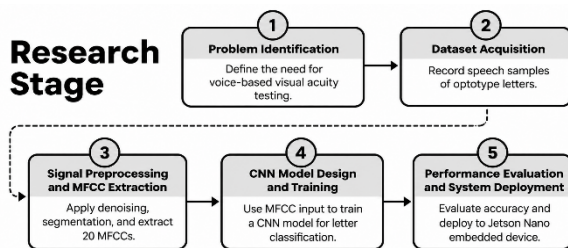


Figure 1. Research Stage

2.1 Overview of the Intelligent Visual Acuity (IVA) System

The proposed Intelligent Visual Acuity (IVA) system integrates voice recognition and deep learning techniques to automate visual acuity assessment. The system mimics the traditional Snellen Chart test but replaces manual interpretation with a machine learning pipeline that captures, processes, and classifies spoken responses using Mel-Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Networks (CNN) [15], [17].

The methodology consists of five major components: dataset collection and augmentation, audio preprocessing, feature extraction using MFCC, model training with CNN architecture, and performance evaluation using a confusion matrix and classification metrics.

2.2 Dataset Collection

The dataset was collected by recording participants' vocal responses to letters displayed on a digital Snellen Chart interface. Participants (N=20) were recruited from Politeknik Negeri Padang campus community under voluntary participation guidelines. All participants provided informed consent prior to data collection. The study involved non-invasive voice recording only, with all data fully anonymized and no personal health or clinical information collected. Therefore, formal ethical approval was not required, in accordance with institutional guidelines for low-risk, non-clinical human subject research. The 20 participants were partitioned at the speaker level into 12 participants for training, 4 for validation, and 4 for

testing. All recordings from each participant were assigned exclusively to one subset, ensuring that no participant appeared simultaneously across the training, validation, and testing sets. Each participant was asked to read aloud letters of decreasing size, simulating a standard VA test. Audio recordings were saved in .wav format and organized into nine subfolders corresponding to each letter used in the Snellen Chart (e.g., C, D, E, F, L, O, P, T, Z), with approximately 200 samples per class [2]. All audio recordings were conducted in a quiet room with consistent microphone placement to ensure signal quality and reproducibility.

To ensure diversity and robustness, recordings were made under various environmental conditions, including background noise and different speaking styles (e.g., fast, slow, whispered). This process followed established practices in speech-based medical diagnostics to enhance generalizability [16], [24]. This approach ensures a speaker-independent evaluation setting, reducing the risk of overestimating model performance.

2.3 Visual Acuity Scoring

Visual acuity scores are derived using standard Snellen chart evaluation rules. During the assessment, participants are instructed to read aloud the optotype letters displayed on a digital Snellen chart interface. A line is considered successfully identified when at least three out of five optotypes are correctly recognized, following conventional clinical practice.

The evaluation proceeds sequentially from larger to smaller optotype sizes. The test is terminated when the participant fails to correctly identify the majority of letters within a given line. The smallest line correctly identified is then used to determine the visual acuity value, expressed in Snellen notation (e.g., 20/20, 20/40).

In the proposed IVA system, the primary function is to automatically recognize spoken optotype responses using the trained classification model. These recognized responses are subsequently used to support visual acuity scoring based on the standard evaluation procedure. It is important to note that this study focuses on the speech-based optotype recognition component and does not include direct clinical validation of visual acuity measurements.

2.4 Data Augmentation

To expand the dataset and improve model generalization, several data augmentation techniques were applied:

Four augmentation operations were applied before dataset partitioning: noise injection, pitch shifting, time stretching, and random gain adjustment. Noise injection added white noise; pitch shifting varied fundamental frequency; time stretching varied speaking rate while preserving pitch; and random gain varied signal amplitude. These operations increased the dataset size tenfold.

In the current experimental pipeline, data augmentation was performed before participant-level partitioning. Augmented samples retained their participant identity during subset assignment; nevertheless, partitioning participants before augmentation and restricting augmentation to the training subset would provide a stricter leakage-control protocol.

2.5 Preprocessing and Feature Extraction

Audio preprocessing comprised spectral-gating noise reduction, amplitude normalization to -1 to $+1$, and segmentation into fixed-length clips corresponding to single optotype responses.

MFCCs were selected as the primary audio feature representation due to their proven effectiveness in capturing perceptually relevant aspects of human auditory processing [25], [26]. The MFCC extraction process consisted of several key steps. First, a pre-emphasis filter was applied to enhance high-frequency components, using the formula $p(n) = s(n) - \alpha \cdot s(n-1)$, where α was set to 0.97. This was followed by framing, where the continuous signal was divided into overlapping short-time segments using a frame size of 25 ms and an overlap of 10 ms. Each frame was then windowed using a Hamming function to reduce spectral leakage. In this implementation, audio signals were sampled at 16 kHz with an FFT window size of 2048 samples and a hop length of 512 samples.

Next, a Fast Fourier Transform (FFT) was applied to convert the time-domain frames into frequency-domain representations. The resulting spectrum was then passed through a Mel-scale filter bank to better approximate human auditory perception. The logarithmic outputs of these filters were subsequently compressed using a Discrete Cosine Transform (DCT), producing the final set of MFCC features. Typically, the top 13 to 20 coefficients are retained, and in this implementation, 20 coefficients per frame were used to strike a balance between dimensionality and discriminative power. These MFCC features were stored in JSON format and served as the input representations during model training [27], [28]. Each utterance was thus encoded as a two-dimensional matrix of size $T \times M$, where T denotes the number of time steps and $M = 20$ represents the number of Mel-frequency cepstral coefficients. A total of 20 MFCC coefficients were extracted from each frame.

2.6 Model Architecture and Training

To effectively model the temporal and spectral features of the MFCC representations, a Convolutional Neural Network (CNN) architecture was employed due to its strength in identifying spatial patterns within image-like data such as spectrograms [29]. The model accepted MFCC matrices of shape $T \times M$ as input, where T represents the number of time steps and $M = 20$ denotes the number of Mel-frequency cepstral coefficients per frame.

The CNN architecture comprised two convolutional blocks, each followed by ReLU activation and MaxPooling, to extract hierarchical temporal-spectral patterns from the MFCC matrices. The extracted feature maps were flattened and passed to a fully connected layer with dropout regularization, followed by a nine-unit SoftMax output layer corresponding to the nine optotype classes. Training used categorical cross-entropy loss and the Adam optimizer. The maximum training limit was 500 epochs, with early stopping based on validation performance to reduce overfitting.

Under the early-stopping criterion, training converged after approximately 46 epochs, as illustrated in Figure 2.

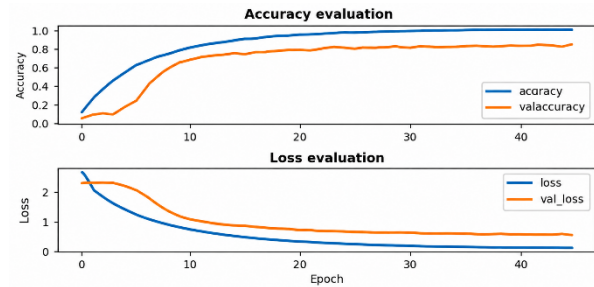


Figure 2. Training and Validation Accuracy and Loss Curves

To ensure speaker-independent evaluation, the dataset was partitioned at the participant level into training, validation, and testing subsets in a 60:20:20 ratio. The 20 participants were allocated to 12 training, 4 validation, and 4 testing participants, with all recordings from each participant restricted to a single subset. Model performance was assessed using accuracy, precision, recall, and F1-score derived from the confusion matrix.

2.7 Evaluation Metrics

The performance of the Intelligent Visual Acuity (IVA) system was assessed using accuracy, precision, recall, and F1-score derived from the confusion matrix. Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy was calculated using Equation 1, while precision and recall were determined using Equations 2 and 3, respectively. The F1-score, representing the harmonic mean of precision and recall, was calculated using Equation 4.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

Label-wise performance varied across optotype classes. Classification accuracy ranged from 97.72% for E to 99.69% for L and O, while precision was lowest for P (84.61%) and reached 100% for O.

2.8 Implementation of Embedded Hardware

The trained model was deployed on an NVIDIA Jetson Orin Nano with a monitor, microphone, and audio card for on-device inference, demonstrating technical feasibility for real-time embedded operation.

The interface guides participant entry, sequential optotype presentation, voice capture, and display of the resulting Snellen-based score.

Having described the full system architecture, including model deployment and user interaction, the following section discusses the experimental results and performance evaluation of the IVA system.

3. Results and Discussions

The dataset collection process involved gathering voice responses from participants reading letters displayed on a digital Snellen Chart interface. The dataset was organized into nine subfolders, each representing a distinct letter (C, D, E, F, L, O, P, T, Z), with approximately 200 audio samples per class. To enhance robustness and generalizability, data augmentation techniques were applied, including noise injection, pitch shifting, time stretching, and random gain adjustment.

These operations increased the dataset size and introduced controlled acoustic variability in noise level, pitch, speaking rate, and signal amplitude. This approach aligns with best practices in speech-based medical diagnostics, where data augmentation improves model resilience against environmental and physiological variations [16], [30].

Following dataset collection and augmentation, raw audio files underwent preprocessing steps including noise reduction, normalization, and segmentation. Mel-Frequency Cepstral Coefficients (MFCC) were extracted as feature vectors, capturing perceptually relevant aspects of human auditory perception [25]. These features were stored in JSON format for model training.

3.1 Model Training and Evaluation

The MFCC-CNN described in Section 2.6 was trained for nine-class spoken-optotype classification using categorical cross-entropy loss and the Adam optimizer. Early stopping terminated training after approximately 46 epochs when validation performance no longer improved.

Evaluation used the participant-level 60:20:20 split described in Section 2.6, corresponding to 12 training, 4 validation, and 4 test participants.

Model evaluation was conducted using a confusion matrix and derived metrics: accuracy, precision, recall, and F1-score. Table 2 presents these results across all classes. As shown in Table 2, class-wise performance was generally high, but the error profile was not uniform. P and E showed the lowest precision values, 84.61% and 86.56%, with recalls of 88.00% and

90.62%, respectively. The simultaneous reduction in precision and recall indicates that the difficulty is not limited to a single type of classification error. Because E and P are short utterances with potentially overlapping temporal-spectral characteristics, pronunciation variation and background noise may produce less separable MFCC representations for these classes. By contrast, O reached 100% precision and 97.50% recall, indicating that the classification difficulty is concentrated in specific acoustically confusable optotypes rather than reflecting uniform degradation across all classes. Future work should examine phoneme-aware features, log-mel representations, or other complementary acoustic representations to improve discrimination between confusable optotypes.

Table 2. Classification Performance for Each Optotype Class

Label	Accuracy	Precision	Recall
C	98.48%	95.60%	93.54%
D	98.78%	94.38%	96.55%
E	97.72%	86.56%	90.62%
F	98.78%	98.66%	91.35%
L	99.69%	98.59%	98.59%
O	99.69%	100%	97.50%
P	97.87%	84.61%	88%
T	98.32%	87.93%	92.72%
Z	99.08%	96.10%	96.10%

Figure 2 shows the training and validation accuracy and loss curves. The training process converged after approximately 46 epochs under the early-stopping criterion.

Table 3. Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1-score
MFCC + kNN	91.82%	89.75%	88.60%	89.17%
MFCC + Random Forest	93.46%	91.24%	90.37%	90.80%
MFCC + SVM	94.73%	92.51%	92.04%	92.27%
MFCC + CNN (Proposed)	97.42%	94.89%	94.33%	94.61%

The results in Table 3 demonstrate that the proposed CNN-based approach outperforms conventional machine learning classifiers across all evaluation metrics. While traditional algorithms such as kNN, Random Forest, and SVM achieve competitive performance when trained on MFCC features, the CNN model consistently achieves higher classification accuracy.

One plausible explanation for the improvement is that the CNN operates directly on structured MFCC feature maps and can learn local temporal-spectral patterns, whereas the conventional baselines use flattened feature vectors. The present comparison supports the selected architecture under the evaluated dataset and protocol but does not establish universal superiority across speech-recognition tasks.

Among the baseline models, SVM achieved the strongest performance, indicating that MFCC features provide strong discriminative capability for spoken optotype recognition. However, the CNN model still demonstrated superior performance, confirming the effectiveness of deep learning methods for this task.

These findings suggest that the performance improvement is not solely attributed to MFCC features, but also to the CNN's ability to capture structured temporal-spectral patterns. This highlights the importance of model architecture in speech-based classification tasks involving short and acoustically similar utterances.

3.2 Deployment of Embedded Hardware

The trained model was deployed on an NVIDIA Jetson Orin Nano integrated with a monitor, microphone, and audio card for real-time spoken-optotype inference. The deployed CNN model has a size of approximately 8.3 MB and achieves an average inference latency of approximately 30–40 ms per prediction on the Jetson Orin Nano platform. GPU utilization during inference remained below 30%, indicating efficient operation on embedded hardware. Figures 3–5 illustrate screenshots of the application interface at different stages of testing, demonstrating that the system offers clear navigation and real-time feedback throughout the VA assessment process.

Figure 3. User Data Input Page

Figure 4. Test Page

Figure 5. Test Results

The proposed IVA system does not replace conventional Snellen chart testing but rather complements it by automating the recognition of spoken optotypes during the examination process. This automation can help reduce manual interpretation errors and improve workflow efficiency during visual acuity assessments [3], [31].

Clinical equivalence was not evaluated: no ophthalmologist- or optometrist-administered Snellen/logMAR assessment was used as a reference standard. Accordingly, the reported results should be interpreted as spoken-optotype classification performance rather than evidence of diagnostic accuracy.

Table 4. Effect of Data Augmentation on Model Performance

Training Configuration	Accuracy	Precision	Recall	F1-score
CNN without augmentation	92.84%	90.31%	89.77%	90.03%
CNN with augmentation (proposed)	97.42%	94.89%	94.33%	94.61%

As shown in Table 4, the CNN achieved 92.84% accuracy without augmentation and 97.42% with augmentation. Because augmentation was applied before partitioning, this difference is reported descriptively and should not be interpreted as a leakage-free estimate of generalization benefit.

3.3 Limitations and Future Work

Several limitations qualify for the reported performance. First, augmentation was applied before participant-level partitioning. Although augmented samples retained participant identity during subset assignment, a stricter protocol would partition speakers before augmentation and augment only the training subset. Second, the dataset contains only 20 participants, limiting external validity across age groups, accents, dialects, and recording environments. The participant-level 12/4/4 split prevents direct speaker overlaps, but stronger unseen-speaker validation using Group K-Fold or LOSO should be prioritized in future work. Third, residual errors are concentrated among acoustically confusable optotypes, particularly E and P, and live predictions remain sensitive to noise and pronunciation variability.

Finally, the study validates spoken-optotype recognition rather than clinical visual acuity accuracy. Clinical comparison with ophthalmologist- or optometrist-administered Snellen/logMAR assessment is required before clinical applicability can be established. Future work should also evaluate larger multilingual cohorts, leakage-safe augmentation, lightweight keyword-spotting models, and more complete embedded profiling (CPU, memory, power, and sustained throughput).

4. Conclusions

This study developed a speech-driven optotype-recognition component for an IVA workflow by combining MFCC features, CNN classification,

Snellen-based scoring logic, and NVIDIA Jetson Orin Nano deployment. The proposed MFCC-CNN achieved 97.42% spoken-optotype classification accuracy and outperformed the evaluated kNN, Random Forest, and SVM baselines under the current protocol. The embedded model occupied approximately 8.3 MB and required about 30–40 ms per prediction. These findings demonstrate technical feasibility for speech-driven optotype recognition within a visual acuity testing workflow; they do not establish clinical equivalence of the resulting visual acuity scores. Future work should use leakage-safe speaker-level evaluation, larger and more diverse cohorts, and formal comparison with clinician-administered Snellen/logMAR assessments.

Declarations

AI Usage Statement

The authors used OpenAI ChatGPT and Grammarly solely to assist with language editing, grammar checking, readability, and formatting during manuscript preparation. These tools were not used to generate scientific content, conduct data analysis, interpret results, or formulate research conclusions. All scientific content, research methodology, data analysis, interpretation of results, and conclusions were developed, reviewed, and verified by the authors, who take full responsibility for the integrity and accuracy of the manuscript.

Author Contributions

D.M: conceptualization, methodology, software, formal analysis, visualization, writing and editing –review and editing. H: supervision, validation and review. F.S: software, formal analysis, visualization. U.I.A: software, formal analysis, visualization. J: software, formal analysis, visualization, writing -original draft.

Funding Statement

This research was supported by the Directorate of Academic Higher Vocational Education through the 2024 *Program Dana Padanan*. The authors gratefully acknowledge the funding provided under Cooperation Agreement Number: 232/PKS/D.D4/PPK.01.APTV/VII/2024 and Number: 238/PL9/KS/2024, dated July 9, 2024.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

This research involved non-invasive voice recordings collected from 20 participants for spoken-optotype recognition experiments. All participants provided informed consent prior to data collection. The recordings were anonymized, and no personal health or clinical information was collected. The authors confirm that the study was conducted in accordance with applicable institutional ethical requirements and that the recordings were used solely for research purposes with participant permission. Under the applicable

institutional guidelines, this low-risk, non-clinical study did not require formal ethical approval. No animal subjects were involved.

Data and Code Availability

The raw voice-recording dataset is not publicly available because speech recordings may contain potentially identifiable biometric information. The dataset was collected with participant consent for research purposes and is subject to privacy and institutional restrictions. Supporting derived data may be made available by the corresponding author upon reasonable request, subject to applicable privacy, participant-consent, and institutional requirements.

References

- [1] A. Fazeenah, "A literary review on visual acuity," *International Journal of Creative Research Thoughts*, vol. 9, no. 6, 2021. <https://ijcrt.org/papers/IJCRT2106378.pdf>
- [2] J. H. Kim *et al.*, "A Deep Learning Ensemble Method to Visual Acuity Measurement Using Fundus Images," *Applied Sciences (Switzerland)*, vol. 12, no. 6, 2022, doi: [10.3390/app12063190](https://doi.org/10.3390/app12063190).
- [3] J. C. Wanten *et al.*, "Evaluation of a visual acuity eHealth tool in patients with cataract," *J. Cataract Refract. Surg.*, vol. 49, no. 3, 2023, doi: [10.1097/j.jcrs.00000000000011088](https://doi.org/10.1097/j.jcrs.00000000000011088).
- [4] V. K. Patil *et al.*, "Converging emotion recognition with AI and IoT," in *Affective Computing Applications using Artificial Intelligence in Healthcare: Methods, approaches and challenges in system design*, 2024, doi: [10.1049/pbhe056e_ch9](https://doi.org/10.1049/pbhe056e_ch9).
- [5] M. E. G. Gendy and M. R. Yuce, "Emerging Technologies Used in Health Management and Efficiency Improvement During Different Contact Tracing Phases Against COVID-19 Pandemic," *IEEE Rev. Biomed. Eng.*, vol. 16, 2023, doi: [10.1109/RBME.2022.3219433](https://doi.org/10.1109/RBME.2022.3219433).
- [6] Evangel Chinyere Anyanwu, Femi Osasona, Opeoluwa Oluwanifemi Akomolafe, Jane Osareme Ogugua, Tolulope Olorunsogo, and Ebere Rosita Daraojimba, "Biomedical engineering advances: A review of innovations in healthcare and patient outcomes," *International Journal of Science and Research Archive*, vol. 11, no. 1, 2024, doi: [10.30574/ijrsra.2024.11.1.0139](https://doi.org/10.30574/ijrsra.2024.11.1.0139).
- [7] Kajal, Dr. Kanchan Joshi, and Dr. Tarannum Shakeel, "Yoga for Young Eyes: Exploring the Effects on Visual Acuity in Adolescents," *International Journal of Latest Technology in Engineering Management & Applied Science*, vol. 14, no. 8, 2025, doi: [10.51583/ijltemas.2025.1408000063](https://doi.org/10.51583/ijltemas.2025.1408000063).
- [8] M. A. Ahulló-Fuster, M. L. Sánchez-Sánchez, E. Varela-Donoso, and T. Ortiz, "Measuring postural stability and health-related quality of life in people with visual impairment: a cross-sectional study," *Physiotherapy*, vol. 123, 2024, doi: [10.1016/j.physio.2024.04.254](https://doi.org/10.1016/j.physio.2024.04.254).
- [9] NCT05414799, "Effect of Exercises on Computer Vision Syndrome," <https://clinicaltrials.gov/show/NCT05414799>, 2022.
- [10] S. Das, L. Kalita, and B. Sarma, "The Impact of Digital Devices on Visual Health: Addressing Digital Eye Strain And Blue Light Exposure," *Journal of Population Therapeutics and Clinical Pharmacology*, 2024, doi: [10.53555/jptep.v3i17.7185](https://doi.org/10.53555/jptep.v3i17.7185).
- [11] S. Fatima, R. W. Siddiqui, and T. W. Siddiqui, "Digital Eye Strain: A Cross-sectional Study Among Undergraduate Medical Students in UAE," *New Emirates Medical Journal*, vol. 5, 2024, doi: [10.2174/0102506882361765250221072850](https://doi.org/10.2174/0102506882361765250221072850).
- [12] S. Bagaji and R. Rao, "Digital Fatigue In The Age of Screens: Eye and Postural Strain Among 18–35-Year-Old Screen Users," *International Journal of Research -GRANTHAALAYAH*, vol. 13, no. 5, 2025, doi: [10.29121/granthaalayah.v13.i5.2025.6191](https://doi.org/10.29121/granthaalayah.v13.i5.2025.6191).
- [13] Y. Gao, C. Ma, L. Guo, X. Zhang, and X. Ji, "CLRD: Collaborative Learning for Retinopathy Detection Using Fundus Images," *Bioengineering*, vol. 10, no. 8, 2023, doi: [10.3390/bioengineering10080978](https://doi.org/10.3390/bioengineering10080978).

- [14] C. Muchibwa, M. H. S. Eldaw, and M. O. Agyeman, "An Assessment of Contemporary Methods and Data-Enabled Approaches for Early Cataract Detection," *F1000Res.*, vol. 12, p. 998, 2024, doi: [10.12688/f1000research.138294.2](https://doi.org/10.12688/f1000research.138294.2).
- [15] S. V. Mahadevkar et al., "A Review on Machine Learning Styles in Computer Vision - Techniques and Future Directions," 2022. doi: [10.1109/ACCESS.2022.3209825](https://doi.org/10.1109/ACCESS.2022.3209825).
- [16] L. D. Mang, F. D. González Martínez, D. Martínez Muñoz, S. García Galán, and R. Cortina, "Classification of Adventitious Sounds Combining Cochleogram and Vision Transformers," *Sensors*, vol. 24, no. 2, 2024, doi: [10.3390/s24020682](https://doi.org/10.3390/s24020682).
- [17] M. Aly and N. S. Alotaibi, "A novel deep learning model to detect COVID-19 based on wavelet features extracted from Mel-scale spectrogram of patients' cough and breathing sounds," *Inform. Med. Unlocked*, vol. 32, 2022, doi: [10.1016/j.imu.2022.101049](https://doi.org/10.1016/j.imu.2022.101049).
- [18] N. Yamsani, A. H. Abdul Hussein, S. J. Faith, K. Al-Attabi, and M. Y. Arabi, "Multimodal Sentimental Analysis Using GloVe Convolutional Neural Network with Capsule Network," in *3rd IEEE International Conference on Mobile Networks and Wireless Communications, ICMNWC 2023*, 2023. doi: [10.1109/ICMNWC60182.2023.10435957](https://doi.org/10.1109/ICMNWC60182.2023.10435957).
- [19] Y. L. Chen, N. C. Wang, J. F. Ciou, S. S. Lien, C. H. Lin, and L. W. Lin, "Speaker recognition using autoencoder-based Mel-frequency cepstral coefficients with convolutional and bidirectional long short-term memory networks," *Smart Science*, 2025, doi: [10.1080/23080477.2025.2565152](https://doi.org/10.1080/23080477.2025.2565152).
- [20] Anitha Mummireddygar and N Ananda Reddy, "Optimizing Speaker Recognition in Complex Environments : An Enhanced Framework with Artificial Neural Networks for Multi-Speaker Settings," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 3, 2024, doi: [10.32628/cseit24103116](https://doi.org/10.32628/cseit24103116).
- [21] D. Subhash, J. L. G., P. B., and V. Ravi, "A robust accent classification system based on variational mode decomposition," *Eng. Appl. Artif. Intell.*, vol. 139, 2025, doi: [10.1016/j.engappai.2024.109512](https://doi.org/10.1016/j.engappai.2024.109512).
- [22] D. J. Min and D. H. Kim, "Speech Emotion Recognition via Sparse Learning-Based Fusion Model," *IEEE Access*, vol. 12, 2024, doi: [10.1109/ACCESS.2024.3506565](https://doi.org/10.1109/ACCESS.2024.3506565).
- [23] P. Wijonarko and A. Zahra, "Spoken language identification on 4 Indonesian local languages using deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, 2022, doi: [10.11591/eei.v11i6.4166](https://doi.org/10.11591/eei.v11i6.4166).
- [24] J. M. van Osta, B. Dreis, E. Meyer, L. F. Grogan, and J. G. Castley, "An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the southern black-throated finch (*Poephila cincta cincta*)," *Ecol. Inform.*, vol. 77, 2023, doi: [10.1016/j.ecoinf.2023.102233](https://doi.org/10.1016/j.ecoinf.2023.102233).
- [25] T. Li, "Optimizing the configuration of deep learning models for music genre classification," *Heliyon*, vol. 10, no. 2, 2024, doi: [10.1016/j.heliyon.2024.e24892](https://doi.org/10.1016/j.heliyon.2024.e24892).
- [26] P. Ruan, X. Zheng, Y. Qiu, and Z. Hao, "A Binaural MFCC-CNN Sound Quality Model of High-Speed Train," *Applied Sciences (Switzerland)*, vol. 12, no. 23, 2022, doi: [10.3390/app122312151](https://doi.org/10.3390/app122312151).
- [27] R. V. Sharan, H. Xiong, and S. Berkovsky, "Benchmarking audio signal representation techniques for classification with convolutional neural networks," *Sensors*, vol. 21, no. 10, 2021, doi: [10.3390/s21103434](https://doi.org/10.3390/s21103434).
- [28] S. J. Park, H. B. Lee, and G. W. Kim, "Eardrum-inspired soft viscoelastic diaphragms for CNN-based speech recognition with audio visualization images," *Sci. Rep.*, vol. 13, no. 1, 2023, doi: [10.1038/s41598-023-33755-2](https://doi.org/10.1038/s41598-023-33755-2).
- [29] H. Li, Q. Yao, and X. Li, "Voiceprint Fault Diagnosis of Converter Transformer under Load Influence Based on Multi-Strategy Improved Mel-Frequency Spectrum Coefficient and Temporal Convolutional Network," *Sensors*, vol. 24, no. 3, 2024, doi: [10.3390/s24030757](https://doi.org/10.3390/s24030757).
- [30] T. Thomas et al., "Modelling dataset bias in machine-learned theories of economic decision-making," *Nat. Hum. Behav.*, vol. 8, no. 4, 2024, doi: [10.1038/s41562-023-01784-6](https://doi.org/10.1038/s41562-023-01784-6).
- [31] C. W. Baker et al., "Comparison of Snellen Visual Acuity Measurements in Retinal Clinical Practice to Electronic ETDRS Protocol Visual Acuity Assessment," *Ophthalmology*, vol. 130, no. 5, 2023, doi: [10.1016/j.ophtha.2022.12.008](https://doi.org/10.1016/j.ophtha.2022.12.008).