Journal of Disability Research

2024 | Volume 3 | Pages: 1–8 | e-location ID: e20240027
DOI: 10.57197/JDR-2024-0027

A Fine-Tuned CatBoost-Based Speech Disorder Detection Model

Ashit Kumar Dutta[1,*] and Abdul Rahaman Wahab Sait[2]

[1]Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Ad Diriyah, Riyadh 13713, Saudi Arabia
[2]Department of Documents and Archive, Center of Documents and Administrative Communication, King Faisal University, Hofuf 31982, Al-Ahsa, Saudi Arabia

Correspondence to:
Ashit Kumar Dutta*, e-mail: adotta@um.edu.sa, Telephone: +966558637380

Received: January 30 2024; Revised: March 7 2024; Accepted: March 12 2024; Published Online: April 2 2024

ABSTRACT

The classification of speech disorders (SDs) is crucial for treating children with speech impairment (SI). An automated SD classification can assist speech therapists in rendering services to children with SI in rural areas. Automated techniques for detecting SDs provide objective assessments of speech attributes, including articulation, fluency, and prosody. Clinical examinations and quantitative assessments provide an in-depth understanding of the patient's speaking abilities and limitations. Existing deep learning (DL) models for SD detection often lack generalization across diverse populations and speech variations, leading to suboptimal performance when applied to individuals with different linguistic backgrounds or dialects. This study introduces a DL-based model for classifying normal and abnormal speeches using voice samples. To overcome the overfitting and bias, the authors construct convolutional neural network models with the weights of MobileNet V3 and EfficientNet B7 models for feature extraction (FE). To improve performance, they integrate the squeeze and excitation block with the MobileNet V3-based FE model. Similarly, the EfficientNet B7-model-based FE is improved using the structure pruning technique. The enhanced CatBoost model differentiates the normal and abnormal speeches using the extracted features. The experimental analysis is performed using the public dataset that contains 4620 utterances of healthy children and 2178 utterances of children with SI. The comparative study reveals the exceptional performance of the proposed SD classification model. The model outperforms the current SD classification models. It can be employed in clinical settings to support speech therapists. Substantial training with diverse voice samples can improve the generalizability of the proposed model.

KEYWORDS

speech disorder, CatBoost, feature engineering, speech therapy, MobileNet V3, EfficientNet B7, speech impairment

INTRODUCTION

Speech disorders (SDs) may affect children's communication and development (Broome et al., 2017). The impact of speech difficulties may differ based on the specific type and extent of the SD (Harding et al., 2013). A significant characteristic of SDs is a lack of fluency in verbal expression of ideas and concepts (Kourkounakis et al., 2021). Children may struggle to speak concisely, build words, or pronounce sounds. SDs may have a negative effect on a child's ability to engage in social relationships (Abaskohi et al., 2022). When children have trouble communicating with their classmates, it may lead to humiliation, frustration, and social isolation. Speech impediment encompasses a wide range of articulation challenges, including problems with generating sounds (Al-Qatab and Mustafa, 2021). Disabilities in speaking may be modest to severe and cover a broad spectrum of medical conditions. Language speech impairment occurs without an underlying mental or physical disease or neurological condition (Bachmann et al., 2021). A language disorder impairs the understanding of oral, written, and additional symbol systems (Booth et al., 2020). Articulation, fluency, and voice abnormalities are the three primary forms of SDs.

Higher-level language processing, understanding, and expression may be affected by neurological diseases (Chaware et al., 2021). Consequently, it may be challenging to construct sentences, understand complex language, and use vocabulary. Neurological conditions may indirectly impact the speech system (Cunningham et al., 2017). Neurological diseases may affect speech output by causing muscular weakness, coordination, or sensory perception adjustments (Jesus et al., 2019). Researchers have employed transdiagnostic and multivariate methods to differentiate neurological disorders (NDs). However, these investigations have not found reliable biomarkers for various NDs. Recent studies have shown that speech characteristics are objective, repeatable, and time-efficient biomarkers for NDs (Laganaro et al., 2021). The classification of SDs can assist in diagnosing NDs. It can identify and track minor changes in the nervous system. Several speech characteristics, including rhythm, fluency, pitch, and articulation, are quantitatively analyzed to classify SDs (Low et al., 2020). Applying these objective metrics allows for a more accurate evaluation of speech features, which may help monitor changes over time. Speech patterns may vary depending on the neurological condition. Based on speech features, classifying SDs may assist in distinguishing Parkinson's, Alzheimer's, Huntington's, and other NDs (Low et al., 2020). Combining SD categorization with additional multimodal measures, including cognitive testing, imaging examinations, and motor function evaluations, could substantially improve the accuracy of diagnosis.

A significant amount of raw speech data are multidimensional and contain numerous distinct data points (Low et al., 2020). Data dimensionality is decreased by identifying and extracting the most relevant information. Feature extraction (FE) is essential for SD classification. It improves classification algorithms' efficiency, interpretability, and generalizability by transforming raw voice data into a more manageable and valuable representation (McFaul et al., 2022). In order to make the classification model more resilient, FE is used to distinguish significant information from irrelevant data and minimize noise. FE enhances computational efficiency during model training and inference by reducing dimensionality. A lower-dimensional feature space allows for more efficient model training, rendering them more scalable and suitable for real-time applications.

Standard and advanced machine learning and deep learning (DL) methods exist to classify SDs (McFaul et al., 2022). Mel-frequency cepstral coefficients are commonly utilized in speech processing (Mogren et al., 2020). Speech signals are classified using their spectral properties. Formants assist in recognizing vowels in speech signals. SDs may be identified by analyzing formant frequencies. Pitch is a fundamental component of prosody in speech. Fluctuations in pitch may serve as an indication of a particular SD (Mogren et al., 2020). Using machine learning techniques allows for extracting key patterns of SDs and their subsequent binary or multiclass categorization (Mugada et al., 2018). Decision trees can classify SDs using categorical data. Convolutional neural networks (CNNs) may learn hierarchical representations of spectrogram images or other speech data for categorization. Recurrent neural networks are appropriate for sequential data. In order to classify disorders, they can detect temporal relationships in voice signals. Long short-term memory (LSTM) can represent long-range relationships in sequential data, making them ideal for SD classification (Nelson et al., 2020). Combining multiple neural network designs, including CNNs and LSTMs, can be beneficial for capturing the spectral and temporal aspects of voice signals.

The capacity of models to generalize can be influenced by the lack of different and well-annotated datasets pertaining to SDs (Pamplona and Ysunza, 2020). It is challenging to build a universal model due to humans' various speech patterns. In order to accommodate individual characteristics, customization is frequently required. Analyzing and interpreting decisions made by DL models is hard. Due to the need for specialized expertise to annotate speech data for individual SD, generating datasets is time-consuming, labor-intensive, and subjective. While developing and implementing SD classification systems, it is necessary to consider several ethical factors, including data privacy and the proper use of sensitive medical data. The study novelties are presented as follows:

- A feature engineering technique for extracting crucial patterns of SDs from the voice samples.
- A fined-tuned CatBoost model for classifying SDs in the resource-constrained environment.

LITERATURE REVIEW

The initial automated speech analysis (ASA) and recognition techniques developed during the 1960s and 1970s could handle isolated sounds from a minimal to moderate vocabulary (Pejovic et al., 2021). Linear predictive coding was devised to accommodate voice tract differences. The ASA tools were improved by technological improvements in the 1980s based on statistical probability modeling, indicating that a specific set of linguistic symbols matched the incoming spoken word signal (Harar et al., 2020). Pediatric SD is characterized by children's inability to produce audible speech. Speech production may be impaired during motor planning, language, or execution (Sisman et al., 2020). Automatic speech analysis has advanced, supporting the use of artificial intelligence (AI) for assessing and treating SD children's speech.

A cost-efficient alternative approach will continue to draw the attention of clients and parents. Nevertheless, current speech evaluation and intervention methods are expensive for children who require intense and long-term speech treatment, preventing effective service delivery. Tappy Talks is an automated technique for diagnosing children's speech apraxia. This technique can detect grouping problems, articulation faults, and prosodic errors. It comprises a clinician interface, a mobile application, and a speech processing engine. Tabby Talks may minimize speech therapists' workload and families' time and resources (Narendra et al., 2021).

Remote speech treatment is becoming prevalent with telepractice and teletherapy (Tracy et al., 2020). Individuals who may not have easy access to in-person services can significantly benefit from this technology (McKechnie et al., 2018). Applications of virtual and augmented reality technology in speech treatment have been investigated. In order to engage individuals in speech exercises and games, they can build immersive and interactive settings. Numerous mobile applications offer speech treatment initiatives. Exercises and collaborative training are standard features of these applications, intended to assist users in improving their language and voice abilities (Usha and Alex, 2023). Voice treatment software helps individuals with voice nodules and vocal cord dysfunction. These applications provide voice workouts and feedback. The physiological components of voice production, including resonance, pitch, and loudness, may be tracked in real time using biofeedback equipment. Receiving rapid feedback helps improve communication. Neuroscientists are investigating the potential of brain–computer interfaces (BCIs) to aid people with profound movement disabilities, such as locked-in syndrome and similar disorders (Mohammed et al., 2020). These interfaces may allow for communication through brain impulses. Transcranial magnetic and direct current stimulation are being studied for their ability to influence neuronal activity and improve speech and language recovery.

Type of speech (isolated words or continuous speech) and lexicon size affect FE, the initial component of ASA techniques (Issa et al., 2020). Thus, the FE and speech acoustic model can influence the accuracy of the ASA tools' performance. Despite the advances in ASA tools, computational modeling systems continue to encounter challenges. Children experiencing developmental growth while having speech problems provide considerably more significant obstacles to ASA speech assessment methods. In order to be considered practical, ASA tools that use diagnostic or therapeutic software have to satisfy reliability requirements identical to those of human raters. McKechnie et al. (2018) state that commonly recognized percentage agreement requirements for perceptual evaluations of speech across two human raters or outcome reliability across two assessments of the same activity are 75-85% (Grill and Tučková, 2016). Despite extensive research on ASR, developing speech treatment tools with ASR capabilities for use in pediatric speech sound problems has received relatively little attention. The automated system demands substantial training with larger speech samples to improve evaluation and treatment feedback.

Existing DL-based SD classification has multiple challenges and knowledge gaps (Grill and Tučková, 2016). The lack of well-designed longitudinal studies that monitor the development of SDs over time is a significant challenge. The ability of DL models to adjust to new speech patterns is crucial for monitoring and intervention purposes. Minimizing the gap between machine learning methods and medical expertise is essential (Grill and Tučková, 2016). Reliable and effective SD categorization requires developers to include domain knowledge and clinical insights (Grill and Tučková, 2016). Many SD models concentrate on distinct categories, making generalization harder. There is a continuous effort to develop models to detect and differentiate different types of SDs. Prioritizing patient outcomes and involvement in model design is paramount. It is essential to ensure that individuals may effectively and advantageously use the technology to meet their specific requirements.

A significant limitation in DL-based SD identification is the lack of annotated datasets that effectively encompass the wide range of speech abnormalities, particularly across various age groups, languages, and cultural backgrounds. Current datasets often include an insufficient number of instances of speech problems that are considered uncommon or less prevalent. This results in biased model training and a diminished capacity to make generalizations. Furthermore, it is essential to establish uniform assessment methodologies and standards to reliably evaluate the effectiveness of speech problem identification models.

RESEARCH METHODOLOGY

The authors use mel-spectrogram (MS) generation, FE, and classification techniques to classify SDs using the voice samples of children. Pretrained CNN models in image classification for SD detection have multiple advantages. Initially, pretrained CNN models, mainly those trained on large image datasets such as ImageNet, have acquired universal visual characteristics capable of extracting pertinent information from input images, such as spectrograms or other representations of voice signals. Utilizing these pretrained models safeguards substantial computing resources and time compared to training a CNN from scratch. In addition, pretrained models have previously undergone rigorous training on a wide range of visual data, resulting in an improved ability to generalize and perform well when applied to tasks involving the identification of SDs. By using fine-tuning or transfer learning approaches, it is possible to customize the parameters of the pretrained model to better suit the unique features of SD images. This process improves the model's performance in tasks related to categorization.

Figure 1 depicts the proposed methodology for voice classification. Short-time Fourier transform (STFT) methods are widely used for analyzing signal changes over time. They generate a spectrogram that represents the time and frequency of voice samples. MobileNet V3 and Efficient B7 models extract valuable features from the complex images. These models require limited computational power for FE. In addition, CatBoost is an efficient gradient-boosting model that uses decision trees for classification. The features of these techniques motivate the authors to apply them in the proposed study.

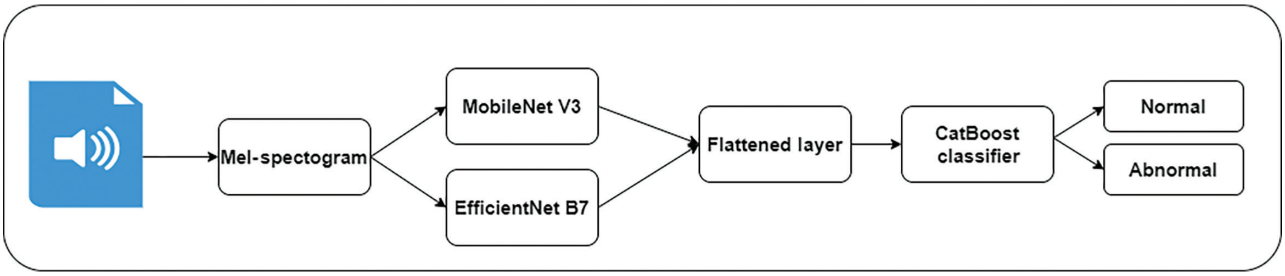

Figure 1: The proposed SD classification model. Abbreviation: SD, speech disorder.

Voice acquisition

The LANNA research group's speech dataset (LANNA), Czech Technical University, Prague, is used to train the proposed model (MobileNet V3; https://github.com/kuan-wang/pytorch-mobilenet-v3). It contains voice samples of 70 healthy children and 33 children with speech impairment (SI), with a total of 4620 utterances of healthy children and 2178 utterances of children with SI. The speech therapists were employed to assess the children. They recorded multiple types of utterances of children aged 4 to 10 years. The normal children's voices were recorded using a Sony digital Dictaphone. The sampling frequency of 16 KHz and 16-bit resolution in the standardized WAV format was followed during the voice recording process. The voice samples of children with SI were recorded in a speech and language therapist's office. The voice recordings were conducted using the sampling frequency of 44 KHz and 16-bit resolution in Mono mode. However, the voice samples contained background noises.

Voice preprocessing

Voice normalization is a frequently used procedure for normalizing audio data prior to MS generation. Normalization standardizes audio signal amplitude to protect retrieved characteristics from signal intensity adjustments. Feature robustness and comparability across recordings may be improved using normalization. The authors employ peak amplitude normalization to scale the signal to a specified value. Root mean square (RMS) normalization is used to scale the RMS value of the signal in order to reduce the background noises. To balance the frequency spectrum, the authors apply a preemphasis filter. Equation 1 presents the mathematical form of the preemphasis filter.

$$Y(t) = X(t) - \propto, \tag{1}$$

where $Y(t)$ is the output voice signal at time (t), X(t) is the input voice signal at t, and $\propto$ is the preemphasis coefficient.

Equation 2 shows the normalized audio of healthy children.

$$N = T \times \frac{H}{M}, \tag{2}$$

where N is the normalized voice, T is the target amplitude, H is the healthy voice, and M is the maximum amplitude.

In order to normalize the voice samples of children with SI, a distortion factor of 0.7 is used. It is used to maintain a consistent peak amplitude. Equation 3 shows the voice normalization of children with SI.

$$N = T \times \frac{D}{M}, \tag{3}$$

where N is the normalized voice, T is the target amplitude, D is the abnormal voice, and M is the maximum amplitude.

The STFT function analyzes signal frequency across time. It shows a signal's frequency content over time. In order to examine the signal across short time frames, the STFT employs a window technique. Equation 4 presents the mathematical form of STFT for MS generation.

$$S(t, f) = \int_{-\infty}^{\infty} S(t) \times W(\tau - t) e^{-j^{2\pi ft}} \,\mathrm{dt}, \tag{4}$$

where $S(t, f)$ is the STFT, $S(t)$ is the voice sample, $W(\tau - t)$ is the window function, f is the frequency, and t is the time.

The authors apply the Hamming window function to divide the voice signals into overlapping segments. Equation 5 presents the mathematical representation of the Hamming window function.

$$W(n) = 0.5 - 0.46\,Cos\left(\frac{2\pi n}{N-1}\right), \tag{5}$$

where n is the input signal, and N is the number of voice samples.

The authors use the Librosa library to generate MS images. The magnitude of the STFT represents the power spectrum of the signal. An Amplitude_to_db function is used to convert the magnitude spectrum to decibels. It is a typical approach to produce values on a logarithmic scale. Using the power spectrum, the MS function generates the mel-scale. It divides the frequencies into mel-filterbanks and calculates the energy by summing the power values. As a result, a matrix form of MS is generated. The visualization function is used to generate the MS images.

Feature extraction

In order to generate diverse features of MS images, the authors construct two CNN models. Figure 2 presents the suggested FE process. The first CNN model with four convolutions, batch normalization, and dropout layers is constructed using the MobileNet V3 model. To improve the model's efficiency, the authors introduce an attention mechanism with squeeze

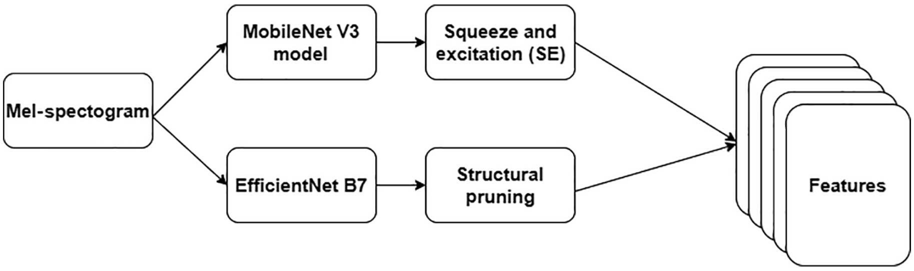

Figure 2: The suggested feature extraction.

and excitation (SE) blocks to extract features adaptively based on their significance.

The MS images are resized into 224 × 224 for FE. Equation 6 highlights the mathematical form of integrating the SE block with the MobileNet V3 model.

$$Features = MobileNet\,V3\_With\_SE(I,F,K,E,S,SR), \quad (6)$$

where I is the image, F is the number of filters, K is the kernel size, E is the expansion ratio, S is the stride for the depthwise separable convolution layers, and SR is the squeeze ratio for the SE block. A flattened layer is used to obtain the features as a vector.

The second FE model applies the weights of the Efficient B7 model. The authors employ four convolutions, batch normalization, and dropout layers to extract the features. To minimize the computational resources, structured pruning is employed. The authors frame the criteria for pruning using the weights and gradients. The pruning process is performed iteratively to refine the model's performance. The extracted features are forwarded to the flattened layer.

Voice classification

CatBoost is a gradient-boosting variant that operates on numerical and categorical features. Categorical features are transformed into numerical ones without feature encoding approaches, including a one-shot or label encoder. Additionally, it employs a symmetric weighted quantile sketch to automatically manage missing values to avoid overfitting and increase dataset performance. However, the performance of the CatBoost model may be reduced in the large dataset. Consequently, generalizing the proposed model in the real-time setting may be challenging. Thus, the author applies Hyperband optimization for tuning the CatBoost parameters, including learning rate, depth, iterations, and leaf number. To ensure the significance of the categorical features, they introduce the SHapley Additive exPlanation (SHAP) analysis. SHAP measures the importance of extracted features. Equations 7 and 8 show the SHAP-integrated CatBoost model.

$$model = Hyperband\;(CatBoost(I,D,L,Loss)), \quad (7)$$

where I is the iteration, D is the depth of the Tree, L is the learning rate, and $Loss$ is the log loss function.

$$SHAP_{value} = SHAP.Explainer\;(model), \quad (8)$$

where the $SHAP_{value}$ is the significant features and *SHAP.Explainer()* is the function that integrates the CatBoost model with the SHAP analysis model.

Using the $SHAP_{value}$, the significant features can be visualized. As a result, the CatBoost classifier performance can be improved by repeating the Hyperband optimization with the essential features.

Evaluation metrics

A model's ability to discriminate between normal speech and SDs is measured using multiple evaluation metrics. The accuracy of a classifier is a measure of how well the proposed SD classifier performs across all classifications. It is appropriate for the balanced dataset. Precision measures the positive prediction accuracy of the proposed model. It is the ratio of accurately predicted positives to total expected positives. Recall indicates how well the proposed model can identify positive instances. It is the percentage of valid positive observations divided by the total number of positive predictions. The F1 score is calculated by determining the harmonic mean of the recall and precision scores. It offers a trade-off between accuracy and recall and is beneficial when there is a disparity in the distribution of classes. A classifier's specificity evaluates its ability to recognize negative occurrences. It is the ratio of predicted negatives to actual negatives. In addition, the authors employ Cohen's Kappa and compute standard deviation, confidence interval (CI), and prediction loss to evaluate the proposed model's reliability.

RESULTS AND DISCUSSIONS

The experimental analysis was conducted using Windows 10 Professional, i5 processor, 16 GB RAM, and NVIDIA R350X Titan configuration. The details of the computational strategies are listed in Table 1. The dataset is divided into a train set (70%) and a test set (30%). In addition, the authors trained the model using 20 epochs and 18 batches. They implemented the proposed model using PyTorch, Librosa, Keras, and TensorFlow libraries. In order to compare the proposed SD model with existing models, the authors used pretrained MobileNet V3 (https://github.com/kuan-wang/pytorch-mobilenet-v3), DenseNet 201 (https://github.com/topics/densenet-201), and SqueezeNet (https://github.com/forresti/SqueezeNet) models.

The performance of the proposed SD model in finding individual classes is presented in Table 2. To address the challenges in preprocessing the audio files, the authors employed STFT-based MS image generation, voice preprocessing, and fine-tuned CatBoost models. The suggested processes yielded a better outcome. The authors fine-tuned the FE model using the attention mechanism and structure pruning techniques. The results indicate the importance of

Table 1: Computational strategies.

Parameters	Values
Epochs	20
Batches	18
MobileNet V3 learning rate	0.0001
MobileNet V3 dropout ratio	0.2
EfficientNet B7 learning rate	0.0001
EfficientNet B7 dropout ratio	0.2
Regularization	L1 and L2
CatBoost learning rate	0.05
CatBoost iterations	100
CatBoost depth	10
CatBoost loss function	Log loss

Table 2: Findings of the proposed model's multiclass classification.

Classes	Accuracy	Precision	Recall	F1-score	Kappa	Sensitivity
Normal	99.2	98.7	99.1	98.9	97.4	97.8
Abnormal	98.9	98.6	98.8	98.7	97.5	98.3
Average	99.0	98.6	98.9	98.8	97.4	98.0

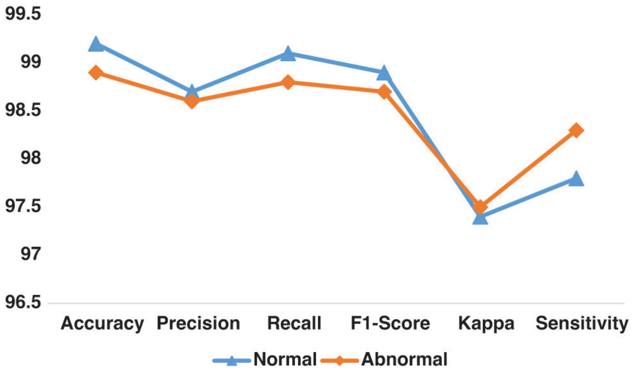

Figure 3: The performance analysis outcomes.

the recommended FE. In addition, the performance of the proposed model is highlighted in Figure 3.

The epoch-wise performance of the suggested SD model is presented in Table 3. The findings highlighted the absence of bias and overfitting in the proposed model. The model yielded an outstanding performance in the specified epochs.

The outcomes of the generalization of the proposed model are presented in Table 4. The proposed SD model overcomes the challenges in transforming the voice samples and extracting the crucial features from the MS images. In addition, the integration of SHAP and CatBoost has assisted the proposed model in delivering an outstanding result. The proposed model has outperformed the current models by scoring with optimal accuracy. Figure 4 shows the models' generalization performance.

Table 5 offers each SD model's computational loss, number of parameters, and floating point operations. The proposed model required less computational power to classify the voice samples than the existing model. It maintained a trade-off between optimal performance and computational resources. In addition, it generated minimal loss and produced reliable results.

This research demonstrates a novel attempt to use DL to categorize SDs. DL is used to classify normal and abnormal children's voices with greater precision. Clinicians may benefit from DL-based SD classification for children for screening, diagnosis, and intervention. Children with speech difficulties may benefit from more efficient and accurate evaluations made possible by incorporating DL-based SD categorization into therapeutic procedures. A child's speech patterns may be objectively evaluated with the use of automated SD classifiers. These patterns can assist physicians in diagnosing specific speech abnormalities faster and better. Remote monitoring of children's speech development may be made possible using the proposed model in the context of teletherapy. Healthcare centers can potentially reach children in rural or underdeveloped communities who lack access to clinical treatments. While the proposed model outcomes are promising, physicians, academics, and technology developers must work together to employ them ethically in clinical settings.

Table 3: Findings of the epoch-wise analysis.

Epochs	Accuracy	Precision	Recall	F1-score	Kappa	Sensitivity
4	97.4	97.5	98.6	98.0	95.7	97.7
12	98.3	97.9	97.9	97.9	96.6	98.1
16	96.7	98.3	98.6	98.4	95.8	98.3
20	99.3	98.9	98.8	98.8	96.7	98.4

Table 4: Outcomes of comparative analysis.

Models	Accuracy	Precision	Recall	F1-score	Kappa	Sensitivity
Proposed SD model	99.0	98.6	98.9	98.8	97.4	98.0
MobileNet V3	96.7	96.4	96.6	96.5	94.3	96.8
DenseNet 201	95.3	95.1	95.3	95.2	95.1	95.4
SqueezeNet	94.8	94.6	94.7	94.6	94.0	94.6
Mohammed et al. (2020)	96.4	96.7	96.6	96.6	96.5	96.2
Issa et al. (2020)	97.3	97.2	97.0	97.1	96.7	96.8

Abbreviation: SD, speech disorder.

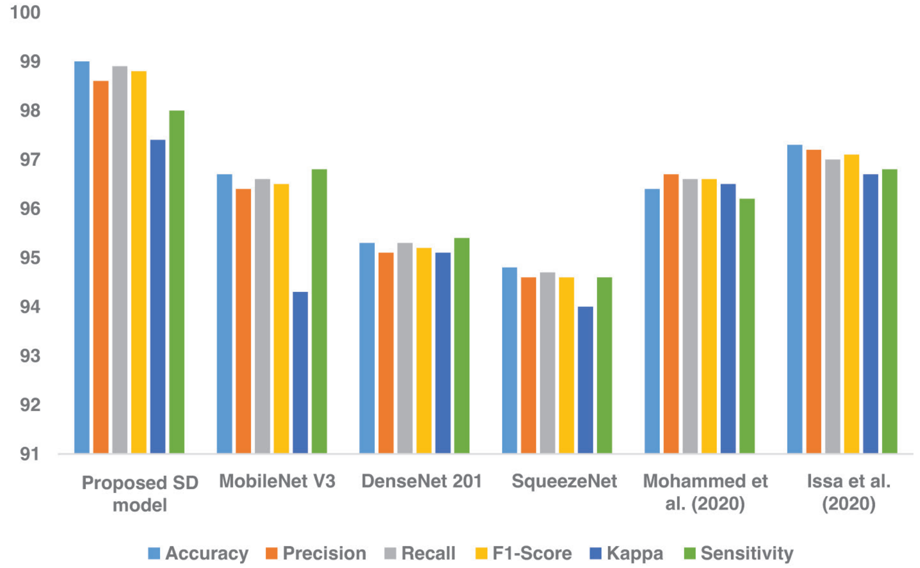

Figure 4: The comparative analysis.

Table 5: Computational strategies and uncertainty analysis.

Models	Standard deviation	CI	Loss	Testing time (in seconds)	Parameters (in millions)	FLOPs (in giga)
Proposed SD model	0.0003	97.7-98.2	1.03	1.25	27	35
MobileNet V3	0.0005	97.2-97.5	1.31	1.34	36	43
DenseNet 201	0.0004	95.8-96.3	1.28	1.45	46	53
SqueezeNet	0.0003	96.1-96.8	1.15	1.98	37	42
Mohammed et al. (2020)	0.0003	95.6-96.3	1.21	1.36	36	58
Issa et al. (2020)	0.0003	95.1-95.9	1.56	1.75	29	47

Abbreviations: CI, confidence interval; FLOPs, floating point operation; SD, speech disorder.

CONCLUSION

In this study, the authors addressed the existing limitations in SD classification. They employed a technique to generate MS images using voice samples. Two FE based on the MobileNet V3 and EfficientNet B7 models were suggested to create diverse features. The authors integrated an attention mechanism with the MobileNet V3-based FE to generate valuable features. They enhanced the EfficientNet B7-based FE using the structure pruning technique. The CatBoost classifier was employed to classify the features into normal and abnormal classes. The SHAP analysis technique was introduced to find the importance of features in improving the CatBoost classifier's performance. The generalization of the proposed model was conducted using the LANNA dataset. The findings revealed the significant performance of the proposed SD model. The model achieved an optimal access of 99.0% and a sensitivity value of 98.0%. It outperformed the existing SD models by overcoming the shortcomings in voice sample classification. The model can be deployed in speech improvement centers. Integrating several data modalities, including spectrograms, facial expressions, and voice recordings, allows for a deeper comprehension of speech challenges. Multimodal models can use additional information from other sources to enhance accuracy and resilience in the detection and diagnosis process. Developing DL models to monitor and track speech abnormalities over an extended period. Regular evaluation of speech patterns and monitoring of symptom changes may help identify problems at an early stage, make individualized modifications to therapy, and track success in rehabilitation. Explainable AI methods will empower doctors to comprehend and have confidence in model predictions, improving decision-making and tailored treatment planning. Extending the proposed model using the vision transformer can yield a better outcome.

ACKNOWLEDGEMENTS

The authors extend their appreciation to the King Salman Center for Disability Research (funder ID: http://dx.doi.org/10.13039/501100019345) for funding this work through Research Group no. KSRG-2023-320.

REFERENCES

Abaskohi A., Mortazavi F. and Moradi H. (2022). Automatic speech recognition for speech assessment of Persian preschool children. arXiv preprint arXiv:2203.12886.

Al-Qatab B.A. and Mustafa M.B. (2021). Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features. *IEEE Access*, 9, 18183-18194. 10.1109/ACCESS.2021.3053335.

Bachmann A.S., Wiltfang J. and Hertrampf K. (2021). Development of the German speech intelligibility index for the treatment of oral cancer patients. *J. Cranio-Maxillofac. Surg.*, 49, 52-58. 10.1016/J.JCMS.2020.11.009.

Booth E., Carns J., Kennington C. and Rafla N. (2020). Evaluating and improving child-directed automatic speech recognition. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association; pp. 6340-6345.

Broome K., McCabe P., Docking K. and Doble M. (2017). A systematic review of speech assessments for children with autism spectrum disorder: recommendations for best practice. *Am. J. Speech Lang. Pathol.*, 26, 1011-1029. 10.1044/2017_AJSLP-16-0014.

Chaware S., Dubey S., Kakatkar V., Jankar A., Pustake S. and Darekar A. (2021). The systematic review and meta-analysis of oral sensory challenges in children and adolescents with autism spectrum disorder. *J. Int. Soc. Prev. Community Dent.*, 11, 469-480. 10.4103/JISPCD.JISPCD_135_21.

Cunningham B.J., Washington K.N., Binns A., Rolfe K., Robertson B. and Rosenbaum P. (2017). Current methods of evaluating speech-language outcomes for preschoolers with communication disorders: a scoping review using the ICF-CY. *J. Speech Lang. Hear. Res.*, 60, 447-464. 10.1044/2016_JSLHR-L-15-0329.

Grill P. and Tučková J. (2016). Speech databases of typical children and children with SLI. *PLoS One*, 11(3), e0150365.

Harar P., Galaz Z., Alonso-Hernandez J.B., Mekyska J., Burget R. and Smekal Z. (2020). Towards robust voice pathology detection: investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases. *Neural Comput. Appl.*, 32, 15747-15757.

Harding S.A., Goldbart J., Morgan L., Parker N., Lewis E., Marshall S., et al. (2013). A systematic review of the interventions used with preschool children with primary speech and language impairment. https://www.researchgate.net/publication/315077019_A_systematic_review_of_the_interventions_used_with_preschool_children_with_primary_speech_and_language_impairment. Accessed June 19, 2022.

Issa D., Demirci M.F. and Yazici A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process Control*, 59, 101894.

Jesus L.M.T., Martinez J., Santos J., Hall A. and Joffe V. (2019). Comparing traditional and tablet-based intervention for children with speech sound disorders: a randomized controlled trial. *J. Speech Lang. Hear. Res.*, 62, 4045-4061. 10.1044/2019_JSLHR-S-18-0301.

Kourkounakis T., Hajavi A. and Etemad A. (2021). Fluentnet: end-to-end detection of stuttered speech disfluencies with deep learning. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 29, 2986-2999.

Laganaro M., Fougeron C., Pernon M., Levêque N., Borel S., Fournet M., et al. (2021). Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in French: the MonPaGe-screening protocol. *Clin. Linguist. Phon.*, 35, 1060-1075. 10.1080/02699206.2020.1865460.

Low D.M., Bentley K.H. and Ghosh S.S. (2020). Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig. Otolaryngol.*, 5, 96. 10.1002/LIO2.354.

McFaul H., Mulgrew L., Smyth J. and Titterington J. (2022). Applying evidence to practice by increasing intensity of intervention for children with severe speech sound disorder: a quality improvement project. *BMJ Open Qual.*, 11, e001761. 10.1136/bmjoq-2021-001761.

McKechnie J., Ahmed B., Gutierrez-Osuna R., Monroe P., McCabe P. and Ballard K.J. (2018). Automated speech analysis tools for children's speech production: a systematic literature review. *Int. J. Speech Lang. Pathol.*, 20, 583-598. 10.1080/17549507.2018.1477991.

Mogren Ã., Sjögreen L., Barr Agholme M. and McAllister A. (2020). Orofacial function in children with speech sound disorders persisting after the age of six years. *Int. J. Speech Lang. Pathol.*, 22, 526-536. 10.1080/17549507.2019.1701081.

Mohammed M.A., Abdulkareem K.H., Mostafa S.A., Abd Ghani M.K., Maashi M.S., Garcia-Zapirain B., et al. (2020). Voice pathology detection and classification using convolutional neural network model. *Appl. Sci.*, 10(11), 3723.

Mugada V., Kiran Kolakota R., Karri S.R., Karineedi T. and Ikram F. (2018). Evaluation of quality of life of head and neck cancer patients: a descriptive cross-sectional study. *Int. J. Res. Rev.*, 5, 241-249.

Narendra N.P., Schuller B. and Alku P. (2021). The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 29, 1925-1936.

Nelson T.L., Mok Z. and Ttofari Eecen K. (2020). Use of transcription when assessing children's speech: Australian speech-language pathologists' practices, challenges, and facilitators. *Folia Phoniatr. Logop.*, 72, 131-142. 10.1159/000503131.

Pamplona M.D.C. and Ysunza P.A. (2020). Speech pathology telepractice for children with cleft palate in the times of COVID-19 pandemic. *Int. J. Pediatr. Otorhinolaryngol.*, 138, 110318. 10.1016/j.ijporl.2020.110318.

Pejovic J., Cruz M., Severino C. and Frota S. (2021). Early visual attention abilities and audiovisual speech processing in 5-7 month-old down syndrome and typically developing infants. *Brain Sci.*, 11, 939. 10.3390/BRAINSCI11070939.

Sisman B., Yamagishi J., King S. and Li H. (2020). An overview of voice conversion and its challenges: from statistical modeling to deep learning. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 29, 132-157.

Tracy J.M., Özkanca Y., Atkins D.C. and Ghomi R.H. (2020). Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *J. Biomed. Inform.*, 104, 103362.

Usha G.P. and Alex J.S.R. (2023). Speech assessment tool methods for speech impaired children: a systematic literature review on the state-of-the-art in speech impairment analysis. *Multimed. Tools Appl.*, 82, 35021-35028.