Transfer Learning-Based Detection of Dysarthric Speech Using Lightweight Convolutional Neural Networks

Henry Ardian Irianta[1*], Abdul Fadlil [2], Rusydi Umar [3]

[1,3] *Master Program of Informatics, University of Ahmad Dahlan, Yogyakarta, Indonesia*
[1] *Department of Information System, University of Siber Muhammadiyah, Yogyakarta, Indonesia*
[2] *Department of Electronic, University of Ahmad Dahlan, Yogyakarta, Indonesia*

*corr-author: henryai@sibermu.ac.id

Abstract - Automatic Speech Recognition (ASR) for a typical speech, such as dysarthria, presents a significant challenge due to high acoustic variability, which often leads to failures in standard models. This challenge is further compounded when implementation is targeted for edge devices with limited computational resources, memory, and power. The need for model architectures that are not only accurate but also highly efficient (lightweight) is crucial for realizing on-device ASR systems with low latency. This research focuses on exploring modern deep learning architectures to address these two primary challenges: accuracy in dysarthric speech and computational efficiency. The study aims to implement and evaluate three efficient models—MobileNetV3Small, EfficientNetB0, and NASNetMobile—on the UASpeech and TORGO datasets. The methodology involves extracting Mel-Frequency Cepstral Coefficients (MFCC) features, which are visualized as spectrograms and subsequently classified using a transfer learning approach. Experimental results show that the MobileNetV3Small model achieved the highest performance on the UASPEECH dataset, attaining a uniform score of 97,8 % for accuracy. This study concludes that lightweight CNN architectures like MobileNetV3Small are highly effective for dysarthric speech classification and demonstrate the feasibility of developing robust and practical ASR systems for resource-constrained environments.

Keywords: automatic speech recognition, dysarthria, lightweight model, transfer learning, MFCC

I. INTRODUCTION

Dysarthria is a motor speech disorder caused by neurological damage that affects articulation, speed, and intonation, degrading quality of life [1]. Early detection and objective classification of its severity using technology are crucial for complementing subjective clinical evaluations [2]. In recent years, deep learning, particularly Convolutional Neural Networks (CNNs), has shown success in speech analysis [3] and dysarthria classification [4]. A significant challenge remains in deploying these models onto resource-constrained platforms. This has driven research into optimizing models for on-device deployment like streaming ASR [5] and keyword spotting [6,7]. However, this progress has largely focused on typical English-speaking populations [8], leaving a gap for non-English speakers with neuromotor disabilities, to address this, studies are exploring deployable solutions for pathological speech, such as the ***CapisciAMe*** mobile app for Italian dysarthric speech [9,10], web platforms for Parkinson's assessment [11], and mobile therapy apps [12]. These efforts underscore the urgency for models that are both accurate and computationally efficient for real-time monitoring on portable devices [13]. In line with this, this research conducts a comparative analysis of three efficient CNN architectures: MobileNetV3Small, EfficientNetB0, and NASNetMobile. Their performance will be evaluated on dysarthric speech classification using the UASpeech and TORGO datasets. This study is expected to provide insights into the most effective and efficient models for practical applications in dysarthria detection, continuing the efforts of previous researchers [2,3].

The field has seen significant progress, with reviews noting that deep learning methods like CNNs consistently outperform classic machine learning, achieving accuracies above 95% where traditional methods fall in the 80-90% range [14]. This high performance is exemplified by transformer-based architectures with transfer learning, which have achieved classification accuracy as high as 98.6% on the UASpeech dataset [15].

Alongside the push for accuracy, efficiency for on-device deployment has become a critical research area. This includes the development of portable TinyML systems on platforms like the ESP-32, which demonstrated over 90% accuracy for Indonesian vowel recognition, highlighting the potential for efficient

rehabilitation devices [16]. Other approaches have focused on lightweight feature and model combinations, such as applying MFCC features with an MLP, which reported 100% accuracy in detecting dysarthria in children [17]. This result, while high, may also suggest potential overfitting, presenting a phenomenon for further study on model generalization. Further optimization efforts have introduced techniques like Network Candidate Search (NCS) to scale models like EfficientNet-B0, resulting in architectures with fewer parameters and higher accuracy than previous state-of-the-art models [18].

Beyond just accuracy and efficiency, research has also ventured into improving model interpretability and design. One study introduced the Interpretable Multi-band Feature Extraction Network (IMBFN), which achieved a strong F1-score of 0.8491 in blind clinical tests, demonstrating robust generalization [19]. Other innovations include hybrid models, such as a CNN-SVM combination for recognizing numbers from dysarthria in speech, which improved accuracy by 7.5% over a standard CNN [20]. New pre-training strategies like ccc-wav2vec 2.0 have also been presented, offering significant relative WER improvements by enhancing robustness during pre-training [21].

Although this collective progress is significant, specific research gaps persist. For instance, a systematic comparative analysis of modern lightweight architectures—specifically MobileNetVSmall, EfficientNetB0, and NASNetMobile—applied to the task of dysarthric speech classification across datasets (UASpeech and TORGO) is currently lacking. Furthermore, uncertainty remains regarding the performance and architectural suitability of these particular models, which were originally designed for computer vision, when applied to the unique and complex acoustic patterns of dysarthric speech. Therefore, this study aims to fill these gaps by directly evaluating the trade-off between accuracy and computational efficiency of these models. The results of this research are expected to provide important insights for the development of practical, accurate, and accessible diagnostic aids for dysarthria via portable devices.

II. METHOD

This study proposes and implements a systematic methodology for dysarthric speech classification using a deep learning approach, with a workflow documented in a series of experimental notebooks. The flowchart summarizing the entire research methodology is presented in Fig. 1. The process begins with data acquisition, where two public datasets, UASpeech [22] and TORGO [23], are automatically downloaded and extracted. The next stage is initial audio analysis, which involves visualizing waveplots to gain a qualitative understanding of the audio signals. The core of the preprocessing is feature extraction, where each audio signal is converted into a *Mel-Frequency Cepstral Coefficients* (MFCC) representation with 40 coefficients which is then standardized to a uniform length of 174 frames through padding and truncating techniques [24].

Before training, an architecture analysis is performed on the three selected *Convolutional Neural Network* (CNN) models. The choice of these models is based on the urgency of using lightweight architectures in terms of parameter count and model size, which is highly relevant for potential implementation on edge devices in the future. The evaluated models—MobileNetV3Small [25], EfficientNetB0 [26], and NASNetMobile [27] are popular and efficient architectures, all three of which were designed by Google. This analysis aims to compare the computational complexity (FLOPs) and parameter counts of the three models. The main training and testing scenario utilizes a *transfer learning* approach; each base model pre-trained on ImageNet is frozen, and an identical custom head is added, consisting of GlobalAveragePooling2D, Dense (128, activation='relu'), Dropout (0.5), and an output Dense (2, activation='softmax') layer. Training is conducted for 20 epochs using the Adam optimizer. To ensure a reliable evaluation and prevent data leakage between speakers, a *Group K-Fold* cross-validation strategy is applied [28,29], with val_loss as the metric for saving the best model. Finally, in the results documentation stage, the performance of each experiment is comprehensively visualized through learning curves, classification reports, Under Area Receiver Operating Characteristic (AUROC) and Under Area The Precision-Recall Curves (AUPRC).

A. Dataset Acquisition

The UASpeech dataset [22] features recordings from 8 dysarthric speakers alongside control speakers who recorded the same utterances. In this study, the control subjects are marked with a "C" (e.g., "CM01," "CF04"). Each subject recorded 765 isolated words, including common words like digits (e.g., ‘zero’) and radio alphabet letters (e.g., ‘Alpha’), and uncommon words (e.g., ‘naturalization’) for phonetic diversity.

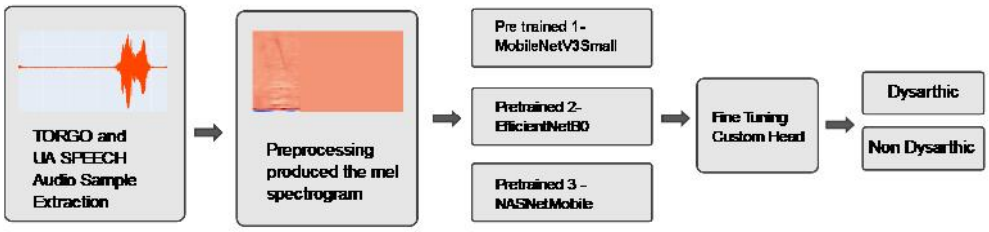


Fig. 1 The proposed comparison model for the detection of dpeech dysarthria

The TORGO dataset was developed to study dysarthric speech from individuals with neuromotor disorders like cerebral palsy [23]. This study uses its "short words" category, which is ideal for ASR development. This category contains a diverse word set, including command terms ("yes," "no"), radio alphabet letters, digits, and words from standardized assessments like the Frenchay Dysarthria Assessment and the Yorkston-Beelman Assessment [23].

The speaker and total data counts presented in Table I are the result of a selection from the entire available data in both datasets. This data subset selection was intentionally made to construct a lightweight modeling experiment scenario. This approach is relevant to the research goal of evaluating efficient models that could potentially be implemented on edge devices with limited computational resources, where training time and dataset size are important considerations, both datasets have uniform audio parameters, which simplifies the preprocessing stage and ensures consistent input to the models. A sample rate of 16000 Hz is a common standard for speech recognition applications, providing a good balance between vocal frequency quality and computational efficiency. A 16-bit depth provides an adequate dynamic range to capture speech nuances, while the use of a single channel (Mono) halves the data load compared to stereo, which is highly suitable for a lightweight modeling scenario. This uniformity allows for the application of the same feature extraction workflow to both datasets without requiring additional resampling or normalization processes.

B. Mel-Spectrogram

Initial analysis of the raw audio signals was conducted to gain a qualitative understanding of the differences between healthy and dysarthric speech. Waveplot visualization was used for this purpose, as it can represent the amplitude of the audio signal over time. For the UASpeech dataset, a pair-matching mechanism was used to compare the same spoken word between dysarthric and control speakers, while for TORGO, random samples from each class were selected for comparison. This initial analysis validates the existence of distinguishable characteristics that can be extracted in the next stage. Fig. 2 presents a comparison of waveplot samples from both datasets.

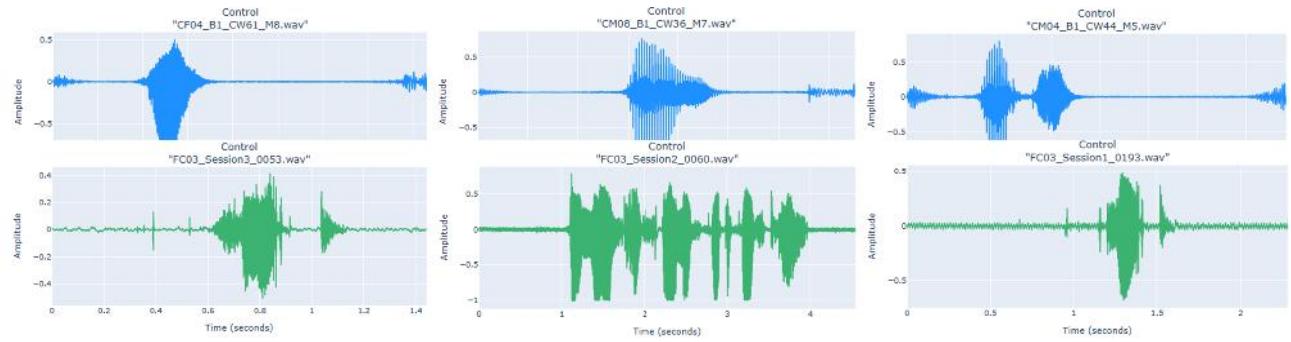
Fig. 2 illustrates significant qualitative differences between the audio signals of control and dysarthric speakers. A comparative analysis of the control samples reveals distinct patterns between the datasets; the UASpeech control samples (blue waveplot top row) generally exhibit compact, high-energy waveforms with clear articulation and short durations, whereas the TORGO control samples (green waveplot bottom row) show more structural variability and longer durations, though still maintaining clear energy peaks. In contrast, the dysarthric samples from both datasets (b) consistently exhibit characteristics associated with motor speech disorders. When compared to their respective control counterparts, the dysarthric signals show significantly longer utterance durations, lower overall amplitude, and more erratic, uncontrolled amplitude fluctuations. The lack of clear silent periods and less defined energy peaks in the dysarthric samples further suggests imprecise articulation across both datasets. Consistent with acoustic literature, dysarthric speech exhibits markers like increased duration and irregular amplitude, which reflect a lack of motor control [30]. To capture these distinguishing characteristics, Mel-Frequency Cepstral Coefficients (MFCCs) were used. The process involved extracting 40 MFCC coefficients, which were then padded or truncated to a uniform length of 174 frames. This created a 40x174 feature matrix that was stacked three times to produce a 3-channel input tensor (40, 174, 3) suitable for the CNN models. TORGO Datasets (Bottom) As visualized in Fig. 3, the MFCC features show clear differences between audio samples. The control samples (a) display energy that is highly concentrated, indicating clear and efficient articulation. In stark contrast, the dysarthric samples (b) show energy that is much more diffuse and "smeared" across both time and coefficients. This distinct visual pattern confirms that MFCCs effectively capture the key features of dysarthric speech, making them an excellent input for the classification models.

TABLE I
SUMMARY OF SPEECH DATASET QUANTITY PER CLASS

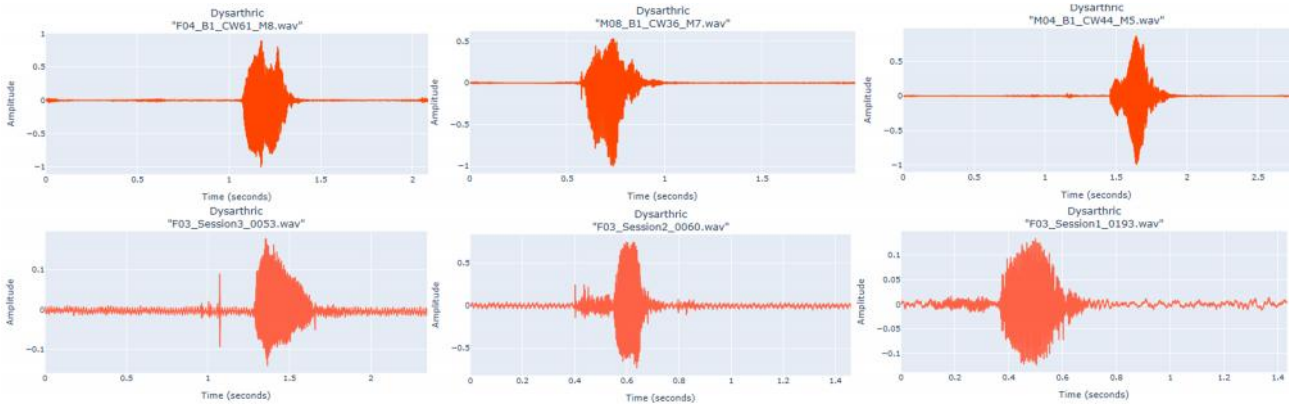
Dataset	Class	Speaker	Total
UASpeech	Disartria	8	5,600
UASpeech	Non-Dysarthria	8	5,610
UASpeech	Total	16	11,210
TORGO	Disartria	8	1,000
TORGO	Non-Dysarthria	7	1,000
TORGO	Total	15	2,000
(UASpeech + TORGO)	Disartria	16	6,600
(UASpeech + TORGO)	Non-Dysarthria	15	6,610
(UASpeech + TORGO)	Total	31	13,210

C. *MobilenetV3, EfficientNetB0, & NASnetMobile Based Feature Extraction*

This study utilizes three modern CNN architectures known for their computational efficiency: MobileNetV3 [25], EfficientNet [26], and NASNet [27]. MobileNetV3 is designed for resource-constrained devices, using depthwise separable convolutions and Squeeze-and-Excitation (SE) modules for high performance in tasks like edge audio processing [31]. EfficientNet achieves superior accuracy and efficiency through a method called compound scaling, which systematically balances network depth, width, and resolution, making it highly valued in the audio domain [38]. NASNet is an architecture automatically discovered via Neural Architecture Search (NAS), resulting in highly efficient building blocks that can outperform manually designed models [27].

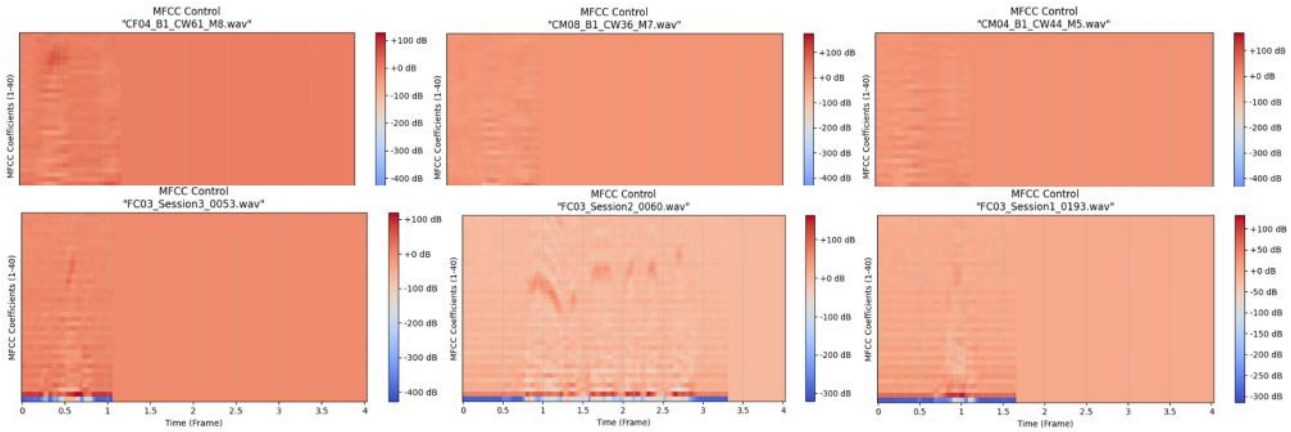


(a)

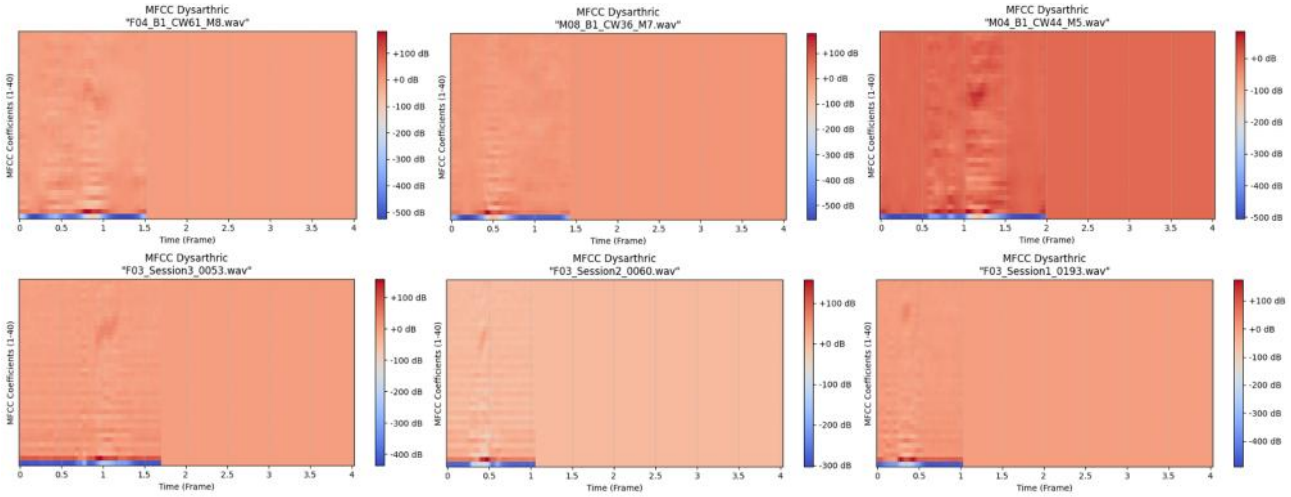


(b)

Fig. 2 Waveplots of audio samples from UASpeech and TORGO datasets: (a) Non-Dysarthric; (b) Dysarthric



(a)



(b)

Fig. 3 MFCC spectrograms of audio samples from the UASpeech and TORGO datasets: (a) Non-Dysarthric; (b) dysarthric

In the experimental setup, the MobileNetV3Small, EfficientNetB0, and NASNetMobile variants were used as base models. For each architecture, the convolutional layers were frozen to retain their pre-trained features from ImageNet. A single, identical custom classification head was added on top of each frozen base. This head is composed of a GlobalAveragePooling2D layer to summarize features, a Dense layer with 128 neurons and ReLU activation, a Dropout layer with a rate of 0.5 for regularization, and a final Dense layer with softmax activation. This structure performs the final binary classification, producing probabilities for the "Dysarthric" and "Non-dysarthric" classes based on the features extracted by each base model.

D. Evaluation Metrics

Parameters and computational configurations used in this experiment are summarized in Table II to ensure reproducibility. Model evaluation uses two metric types: architectural efficiency and classification performance. Efficiency metrics, measured before training for edge device suitability, include **Total Parameters**, **FLOPs**, and **Model Size**. After training, performance is assessed with a **Classification Report** (Precision, Recall, F1-Score), **Learning Curves** to check for overfitting, and **ROC/PRC curves** with Area Under the Curve (**AUC)** values to measure class discrimination.

TABLE II
COMPUTATIONAL CONFIGURATION

Parameter	Value
Feature Extraction	
Number of Coefficients	40
Sequence Length	174 frame
FFT Size	2048
Hop Size	512
Window Function	Hann
Model Architecture: MobileNetV3Small	
Key Architectural Features	Inverted Residual Bottleneck, Depthwise Separable Conv
Main Internal Activation	h-swish
Total Parameters	1.013.234
Trainable Parameters	74.114
Model Size (32-bit)	3.87 MB
Model Architecture: EfficientNetB0	
Key Architectural Features	MBConv Block with Squeeze-and-Excitation
Main Internal Activation	Swish
Total Parameters	4.213.797
Trainable Parameters	164.226
Model Size (32-bit)	16.07 MB
Model Architecture: NASNetMobile	
Key Architectural Features	Normal & Reduction Cells (Normal & Reduction Cells (found via NAS)
Main Internal Activation	ReLU
Total Parameters	4.405.270
Trainable Parameters	135.554
Model Size (32-bit)	16.80 MB
Custom Head Architecture	GlobalAveragePooling2D -> Dense(128, ReLU) -> Dropout(0.5) -> Dense(2, Softmax)
Optimizer	Adam
Loss Function	sparse_categorical_crossentropy
Number of Epochs	40
Batch Size	32
Validation Strategy	Group K-Fold (based on speakers)

III. RESULT AND DISCUSSION

Performance evaluation of the model's was employed by a five-fold *Group K-Fold* cross-validation strategy. Each dataset was systematically partitioned into five subsets based on speaker groups. In each iteration, four sets were used for model training, while the remaining set—containing speakers never seen by the model—was used for testing. This approach guarantees that the model's generalization capability is assessed on entirely new data, maintaining a clinically relevant evaluation framework by preventing data leakage. Furthermore, the use of two different datasets, UASpeech and TORGO, was intended to test the model's generalization capability across diverse data domains. The performance evaluation results are presented in the following section, detailing the average accuracy, precision, recall, and F1-score metrics for the speech sample classification.

TABLE III
FINDINGS OF PERFORMANCE EVALUATION– UA SPEECH DATASET

Model	Accuracy	Precision	Recall	F1-Score
MobileNetV3Small	0.978	0.978	0.978	0.978
EfficientNetB0	0.534	0.759	0.534	0.405
NASNetMobile	0.956	0.956	0.956	0.956

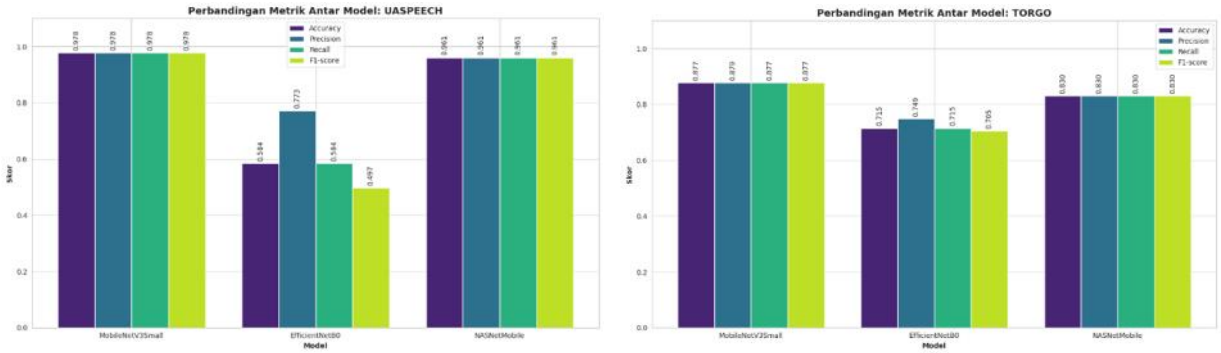
TABLE IV
FINDINGS OF PERFORMANCE EVALUATION– TORGO SPEECH DATASET

Model	Accuracy	Precision	Recall	F1-Score
MobileNetV3Small	0.878	0.879	0.878	0.877
EfficientNetB0	0.715	0.749	0.715	0.705
NASNetMobile	0.830	0.830	0.830	0.830

Based on the performance evaluation, the three models showed varied results. On the UASpeech dataset, as shown in Table III, MobileNetV3Small and NASNetMobile excelled, achieving high F1-scores of 0.977 and 0.956, respectively. In contrast, EfficientNetB0 struggled, scoring only 0.405. On the smaller TORGO dataset, detailed in Table IV, MobileNetV3Small again led with an F1-score of 0.877, followed by NASNetMobile at 0.830. The significant disparity in dataset size between UASpeech (~11,000 samples) and TORGO (~2,000 samples) clearly impacted these outcomes.

On the larger UASpeech dataset, MobileNetV3Small and NASNetMobile were able to leverage the data volume to achieve high and uniform metric scores. However, as vividly illustrated in Fig. 4(a), an anomaly occurred with EfficientNetB0, where the precision metric (0.773) far exceeded the recall (0.584) and F1-score (0.497), indicating a failure to generalize. This is particularly interesting when compared to its performance on the smaller TORGO dataset, as shown in Fig. 4(b), where all models, including EfficientNetB0, showed more "normal" and balanced metric distributions. This phenomenon suggests the inductive bias of a model architecture does not match the data's characteristics. The model might find a "shortcut" or a poor local minimum early in training, and adding more data with similar characteristics could reinforce a model's confidence in this incorrect solution. To address this, several strategies can be applied, such as data augmentation to increase diversity, fine-tuning of several base model layers for better feature adaptation, or adjusting hyperparameters like the learning rate to help the model escape local minimum.

In addition to classification performance metrics, a computational efficiency analysis was also conducted and shown in Table V to evaluate the feasibility of implementing the models on edge devices. The experimental scenario for these metrics was performed theoretically before the training process.

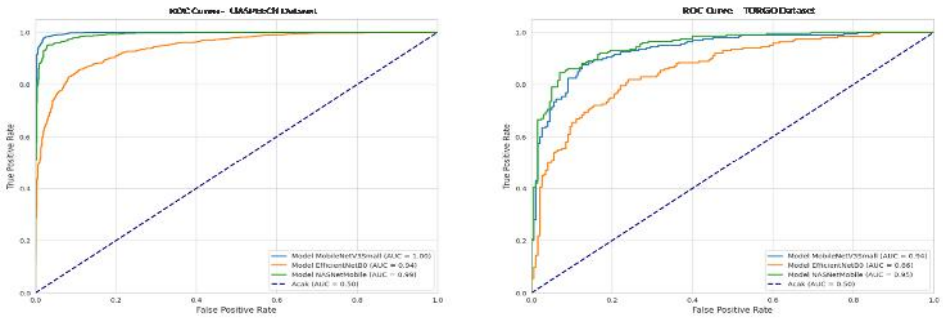


(a) (b)

Fig. 4 Findings of performance evaluation: (a) UASpeech; (b) TORGO

TABLE V
OUTCOMES OF STATISTICAL ANALYSIS AND COMPUTATIONAL EFFICIENCIES

Model	Total Parameter	FLOPs	8-bit Size Estimation	8-bit Activation Memory Estimation
MobileNetV3Small	1,013,234	22.00 MFLOPs	1.09 MB	6.91 KB
EfficientNetB0	4,213,797	148.23 MFLOPs	4.33 MB	15.36 KB
NASNetMobile	4,405,270	202.36MFLOPs	4.85 MB	12.67 KB

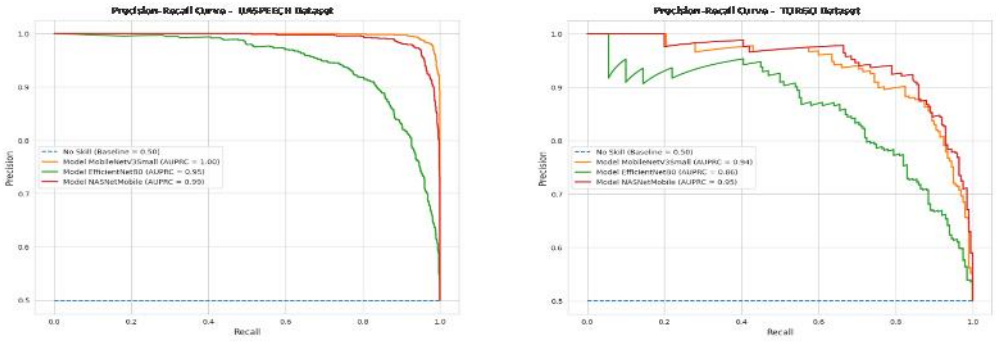


(a) (b)

Fig. 5 AUROC: (a) UASpeech; (b) TORGO datasets

FLOPs were measured by analyzing the computation graph of each model architecture using the TensorFlow profiler to calculate the total number of mathematical operations required for a single inference. The **8-bit Size Estimation** was calculated by saving the 32-bit model to disk, measuring its file size, and then dividing it by four to simulate the effect of quantization. Similarly, the **8-bit Activation Memory Estimation** was calculated by analyzing the output shape of all layers to find the peak RAM requirement, which was then divided by four. As summarized in the experimental results table, **MobileNetV3Small** proved to be the most efficient model, with only **22.00 MFLOPs** and an estimated 8-bit size of **1.09 MB**. **EfficientNetB0** and **NASNetMobile** exhibited a higher computational load, at **148.23 MFLOPs** and **202.36 MFLOPs**, respectively. This analysis confirms that MobileNetV3Small has a significant advantage in terms of efficiency, making it a prime candidate for applications requiring low latency and minimal power consumption.

Further analysis using the ROC and PRC, as shown in Fig. 5 and Fig. 6 respectively, provides a more nuanced view of the models' discriminative abilities. Fig. 5(a) reveal the ROC curves on the **UASpeech dataset** show that MobileNetV3Small achieved a perfect AUC of 1.00, closely followed by NASNetMobile at 0.99. Fig. 5 (b), On the more challenging **TORGO dataset**, NASNetMobile performed best with an AUC of 0.95, slightly outperforming MobileNetV3Small (0.94). The AUPRC in Fig. 6 reinforce these findings. The "No Skill" baseline at 0.50 represents a random classifier in a balanced dataset; any model performing above this line demonstrates skill. Fig. 6(a) On UASpeech, MobileNetV3Small (AUPRC=1.00) and NASNetMobile (AUPRC=0.99) demonstrated an exceptional ability to maintain high precision even at high recall levels. Fig. 6(b) show that on TORGO, NASNetMobile (AUPRC=0.95) and MobileNetV3Small (AUPRC=0.94) again showed robust performance, whereas EfficientNetB0 (AUPRC=0.86) exhibited a more noticeable drop in precision as recall increased. Both sets of curves consistently identify MobileNetV3Small and NASNetMobile as the top-performing models. MobileNetV3Small's balance of accuracy and efficiency makes it a strong candidate for future work. Key research opportunities include validating its performance on real-world edge devices, applying advanced quantization to further optimize its footprint, and extending its application to other languages and speech pathologies through transfer learning. These steps could significantly advance the accessibility of speech recognition for diagnostic aids.



(a) (b)

Fig. 6 AUPRC: (a) UASpeech; (b) TORGO datasets

IV. CONCLUSION

This study successfully evaluated three CNN architectures—MobileNetV3Small, EfficientNetB0, and NASNetMobile—for dysarthric speech classification on the UASpeech and TORGO datasets. The results consistently show that a transfer learning approach is effective, with MobileNetV3Small emerging as the most promising model. It achieved the highest F1-scores (0.977 on UASpeech and 0.877 on TORGO) and proved to be the most efficient with a computational load of only 22.00 MFLOPs, making it an ideal candidate for edge devices.The experiments also revealed that EfficientNetB0 failed to generalize on the larger UASpeech dataset, suggesting a data-architecture mismatch. Furthermore, the high discriminative ability of the models, particularly MobileNetV3Small's near-perfect AUC of 1.00 on UASpeech, indicates the dataset may pose a limited challenge for these powerful architectures. Future research should focus on data augmentation to improve generalization and explore advanced training strategies like fine-tuning, hyperparameter adjustment, or model optimization. Testing these models on more diverse datasets is also crucial to validate their feasibility as an assistive technology for individuals with dysarthria.

REFERENCES

[1] R. Chiaramonte and M. Vecchio, “A Systematic Review of Measures of Dysarthria Severity in Stroke Patients,” *PM R*, vol. 13, no. 3, pp. 314–324, 2021, doi: 10.1002/pmrj.12469.

[2] Q. Miao, M. Zhang, J. Cao, and S. Q. Xie, “Reviewing high-level control techniques on robot-assisted upper-limb rehabilitation,” *Adv. Robot.*, vol. 32, no. 24, pp. 1253–1268, 2018, doi: 10.1080/01691864.2018.1546617.

[3] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, “A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders,” *J. Voice*, vol. 33, no. 6, pp. 947.e11-947.e33, 2019, doi: 10.1016/j.jvoice.2018.07.014.

[4] S. R. Mani Sekhar, G. Kashyap, A. Bhansali, A. Andrew Abishek, and K. Singh, “Dysarthric-speech detection using transfer learning with convolutional neural networks,” *ICT Express*, vol. 8, no. 1, pp. 61–64, 2022, doi: 10.1016/j.icte.2021.07.004.

[5] L. Ben Letaifa and J. L. Rouas, “Transformer Model Compression for End-to-End Speech Recognition on Mobile Devices,” *Eur. Signal Process. Conf.*, vol. 2022-Augus, pp. 439–443, 2022, doi: 10.23919/eusipco55093.2022.9909765.

[6] Y. Zhang, N. Suda, L. Lai, and V. Chandra, “Hello Edge: Keyword Spotting on Microcontrollers,” pp. 1–14, 2017, [Online]. Available: http://arxiv.org/abs/1711.07128

[7] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, “Streaming keyword spotting on mobile devices,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 2277–2281, 2020, doi: 10.21437/Interspeech.2020-1003.

[8] A. Huang, K. Hall, C. Watson, and S. R. Shahamiri, “A review of automated intelligibility assessment for dysarthric speakers,” *2021 11th Int. Conf. Speech Technol. Human-Computer Dialogue, SpeD 2021*, pp. 19–24, 2021, doi: 10.1109/SpeD53181.2021.9587400.

[9] D. Mulfari and M. Villari, “A Voice User Interface on the Edge for People with Speech Impairments,” *Electron.*, vol. 13, no. 7, 2024, doi: 10.3390/electronics13071389.

[10] D. Mulfari, L. Carnevale, and M. Villari, “Toward a lightweight ASR solution for atypical speech on the edge,” *Futur. Gener. Comput. Syst.*, vol. 149, pp. 455–463, 2023, doi: 10.1016/j.future.2023.08.002.

[11] Z. Peng and J. Huang, “Soft rehabilitation and nursing-care robots: A review and future outlook,” *Appl. Sci.*, vol. 9, no. 15, 2019, doi: 10.3390/app9153102.

[12] I. Díaz, J. Catalán, F. Badesa, X. Justo, L. Lledó, A. Ugartemendía, J. J. Gil, J. Díez, N. García-Aracil, "Development of a robotic device for post-stroke home tele-rehabilitation", Advances in Mechanical Engineering, vol. 10, no. 1, 2018.

https://doi.org/10.1177/1687814017752302.

[13] P. Mittal, *A comprehensive survey of deep learning-based lightweight object detection models for edge devices*, vol. 57, no. 9. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10877-1.

[14] Z. Qian and K. Xiao, "A Survey of Automatic Speech Recognition for Dysarthric Speech," *Electron.*, vol. 12, no. 20, pp. 1–23, 2023, doi: 10.3390/electronics12204278.

[15] U. Irshad, R. Mahum, I. Ganiyu, F. Butt, L. Hidri, T. Ali, A. M. El Sherbeeny, "Utran-dsr: a novel transformer-based model using feature enhancement for dysarthric speech recognition", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, no. 1, 2024. https://doi.org/10.1186/s13636-024-00368-0.

[16] B. Riyanta, H. A. Irianta, and B. P. Kamiel, "Development of Speech Command Control Based TinyML System for Post-Stroke Dysarthria Therapy Device," *J. Robot. Control*, vol. 4, no. 4, pp. 466–478, 2023, doi: 10.18196/jrc.v4i4.15918.

[17] A. Fadlil, L. Perdana, A. Pujiyanta, Herman, H. I. K. Fathurrahman, and M. M. J. Samodro, "Implementation of Dysarthria Identification Using MFCC and Multilayer Perceptron Algorithm," *SSRG Int. J. Electr. Electron. Eng.*, vol. 12, no. 1, pp. 32–46, 2025, doi: 10.14445/23488379/IJEEE-V12I1P105.

[18] C. C. Wang, C. Te Chiu, and J. Y. Chang, "EfficientNet-eLite: Extremely Lightweight and Efficient CNN Models for Edge Devices by Network Candidate Search," *J. Signal Process. Syst.*, vol. 95, no. 5, pp. 657–669, 2023, doi: 10.1007/s11265-022-01808-w.

[19] D. Zhao, Z. Qiu, Y. Jiang, X. Zhu, X. Zhang, and Z. Tao, "A depthwise separable CNN-based interpretable feature extraction network for automatic pathological voice detection," *Biomed. Signal Process. Control*, vol. 88, no. PB, p. 105624, 2024, doi: 10.1016/j.bspc.2023.105624.

[20] H. Dyoniputri and Afiahayati, "A hybrid convolutional neural network and support vector machine for dysarthria speech classification," *Int. J. Innov. Comput. Inf. Control*, vol. 17, no. 1, pp. 111–123, 2021, doi: 10.24507/ijicic.17.01.111.

[21] V. S. Lodagala, S. Ghosh, and S. Umesh, "CCC-WAV2VEC 2.0: Clustering AIDED Cross Contrastive Self-Supervised Learning of Speech Representations," *2022 IEEE Spok. Lang. Technol. Work. SLT 2022 - Proc.*, pp. 1–8, 2023, doi: 10.1109/SLT54892.2023.10022552.

[22] J. Yu, X. Xie, S. Liu, S. Hu, M. Lam, X. Wu, K. H. Wong, X. Liu, H. Meng, "Development of the CUHK dysarthric speech recognition system for the UA Speech corpus", Interspeech, 2018. https://doi.org/10.21437/interspeech.2018-1541.

[23] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012, doi: 10.1007/s10579-011-9145-0.

[24] E. B. Sudewo, M. Kunta Biddinika, R. Umar, and A. Fadlil, "Evaluating the Impact of Optimizer Hyperparameters on ResNet in Hanacaraka Character Recognition," *Preserv. Digit. Technol. Cult.*, pp. 1–11, 2025, doi: 10.1515/pdtc-2024-0061.

[25] A. Howard, W. Wang, G. Chu, L. Chen, B. Chen, and M. Tan, "Searching for MobileNetV3 Accuracy vs MADDs vs model size," *Int. Conf. Comput. Vis.*, pp. 1314–1324, 2019.

[26] Q. V. Le Mingxing Tan, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks Mingxing," *Can. J. Emerg. Med.*, vol. 15, no. 3, p. 190, 2013.

[27] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, "Zoph_Learning_Transferable_Architectures_CVPR_2018_paper.pdf," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 8697–8710, 2018.

[28] A. Peryanto, A. Yudhana, and R. Umar, "Klasifikasi Citra Menggunakan Convolutional Neural Network dan K Fold Cross Validation," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 45–51, 2020, doi: 10.30871/jaic.v4i1.2017.

[29] I. B. Mahendra, I. M. G. Sunarya, and I. M. A. Wirawan, "Comparison of Multinomial, Bernoulli, and Gaussian Naïve Bayes for Complaint Classification in Pro Denpasar Application", *JUITA*, vol. 13, no. 1, pp. 77–86, Mar. 2025.

[30] A. R. W. Sait, S. Sankaranarayanan, and P. Gouthaman, "Multi-Feature Fusion-Based Speech Disorder Classification Using MobileNetV3-EfficientNetB7, Linformer-Performer, and SHAP-Aware XGBoost," *IEEE Access*, vol. 13, no. May, pp. 83348–83360, 2025, doi: 10.1109/ACCESS.2025.3562232.

[31] A. Wong, M. Famouri, M. Pavlova, and S. Surana, "TinySpeech: Attention Condensers for Deep Speech Recognition Neural Networks on Edge Devices," pp. 1–10, 2020, [Online]. Available: http://arxiv.org/abs/2008.04245