A Comparative Study of Lightweight CNNs for Dysarthria Classification

Daniel Prem ,
Department of Computer Engineering,
Karunya Institute of Technology and
Sciences, (Deemed to be University),
Karunya Nagar,
Coimbatore,Tamil Nadu, India
email- danielprem@karunya.edu.in

Kumudha Raimond,
Department of Computer Engineering,
Karunya Institute of Technology and
Sciences, (Deemed to be University),
Karunya Nagar,
Coimbatore,Tamil Nadu, India
email- kraimond@karunya.edu

Babita Prem,
School of Management,
S.A. College of Arts & Science,
Chennai, India
email- babitaprem@sacas.ac.in

Abstract— Dysarthria, a motor speech disorder that impacts speech clarity, creates a significant barrier to effective communication. This paper compares four lightweight Convolutional Neural Networks (CNNs), MobileNetV3Small, ShuffleNet, SqueezeNet, and EfficientNetB0, for classifying dysarthric speech using spectrogram images. These models are designed for deployment on devices with limited processing capabilities, where rapid inference is critical. The study considers classification accuracy, per-epoch training time, and number of parameters to assess both performance and the computational demands during model development. Results indicate that EfficientNetB0 achieves the highest accuracy (95.42%) but requires a substantially longer training time per epoch (265 s). This high training time implies that EfficientNetB0 is best utilized in settings where model training can be performed on more powerful hardware ahead of deployment, rather than in scenarios demanding frequent on-device retraining. In contrast, MobileNetV3Small processes data in less training time (28 s per epoch) albeit with lower accuracy (90.90%), while ShuffleNet offers a strong balance with 95.05% accuracy and moderate training time (55 s per epoch). Although SqueezeNet has the fewest parameters, its performance exhibits a higher standard deviation across cross-validation folds and a longer training time (420 s per epoch), indicating challenges in maintaining consistent results. These quantitative insights support informed model selection for dysarthric speech recognition(DSR) based on specific priorities, contributing to enhanced accessibility and assistive communication for individuals with speech impairments.

Keywords — Dysarthria, lightweight models, MobileNetV3Small , ShuffleNet, SqueezeNet and EfficientNetB0 , speech recognition, low-power devices.

I. INTRODUCTION

Dysarthria is a motor speech disorder caused by neurological conditions such as cerebral palsy, traumatic brain injury, and stroke, which significantly impair speech articulation and limit effective communication in daily life. Assistive technologies, particularly automatic speech recognition (ASR) systems, offer potential solutions to enhance accessibility for individuals with dysarthria. However, traditional ASR systems encounter considerable challenges due to the high variability inherent in dysarthric speech, often resulting in poor recognition accuracy (Takashima et al., 2020) [11]. Although deep learning-based ASR models have shown improved performance, their substantial computational requirements make real-time deployment on mobile and embedded devices impractical (Kim et al., 2019) [12].

To address these limitations, this study evaluates four lightweight deep learning architectures—MobileNetV3Small, ShuffleNet, SqueezeNet, and EfficientNetB0—for dysarthric speech classification using spectrogram images. These models are selected for their efficient designs that balance model complexity, inference speed, and classification performance, making them particularly suitable for resource-constrained environments. Derived from architectures originally developed for mobile and embedded applications, MobileNetV3Small (Howard et al., 2019) [13] employs Neural Architecture Search (NAS) along with squeeze-and-excitation layers and hard-swish activations to optimize efficiency and accuracy trade-offs. ShuffleNet (Zhang et al., 2018) [14] utilizes pointwise group convolutions and channel shuffle operations to significantly reduce computational complexity while maintaining high classification performance. SqueezeNet (Iandola et al., 2016) [15] achieves AlexNet-level accuracy with 50× fewer parameters by leveraging fire modules to enhance feature representation within a compact network. EfficientNetB0 (Tan and Le, 2019) [16] applies compound scaling to jointly optimize depth, width, and resolution, achieving state-of-the-art accuracy with fewer parameters compared to conventional CNNs.

Several modifications were introduced in this study to adapt these architectures for dysarthric speech classification from spectrogram images. In MobileNetV3Small, the original classification head was replaced with a custom fully connected layer optimized for binary classification, coupled with transfer learning from ImageNet-pretrained weights. ShuffleNet was modified with optimized group convolutions and feature extraction layers to improve dysarthric speech recognition without compromising efficiency. SqueezeNet was enhanced with batch normalization and dropout layers to improve generalization and stabilize training. EfficientNetB0 was tailored with a two-branch architecture incorporating a squeeze-and-excitation (SE) block to refine feature representation, thereby improving classification accuracy.

The results of this study contribute to the development of real-time, energy-efficient ASR systems for individuals with dysarthria. The proposed models are evaluated based on accuracy, precision, recall, specificity, and F1-score to determine their effectiveness for dysarthric speech classification.

The remainder of this paper is organized as follows: Section II provides an in-depth review of the literature, discussing existing approaches, related methodologies, and studies that have addressed the challenges in dysarthria classification. Section III details the proposed methodology, including comprehensive discussions on data collection and preprocessing techniques, the conversion of raw speech signals into spectrogram representations, the selection and modification of CNN architecture models tailored for dysarthric speech, and the training and validation protocols employed. Section IV offers a thorough model implementation analysis, examining key aspects such as training dynamics, computational efficiency, model stability, variance across different data splits, and the precision-recall trade-offs along with specificity, ultimately summarizing the overall performance and accuracy of the models. Section V presents the performance evaluation metrics, detailing quantitative results and comparative insights that underscore the strengths and limitations of each approach. Finally, Section VI concludes the paper with a synthesis of the findings, discusses their implications for real-time assistive communication systems, and provides suggestions for future research directions aimed at further improving dysarthria classification and enhancing system robustness.

II. LITERATURE REVIEW

The field of dysarthric speech classification has seen significant advancements through various methodologies, leveraging deep learning techniques and lightweight model implementations to improve accuracy and efficiency. This section provides an overview of key research initiatives, highlighting the evolution of approaches used in DSR and related speech processing tasks.

Joshy and Rajan (2021)[1] conducted a comprehensive study on deep learning for dysarthric speech classification, employing CNN, DNN, and LSTM models trained on the TORGO and UA-Speech datasets. Their results demonstrated high classification accuracies of 96.18% and 93.24%, respectively, reinforcing the effectiveness of deep learning architectures in handling the complex characteristics of dysarthric speech.

Suhas et al. (2020)[2] used log Mel spectrograms with CNNs for dysarthric speech classification on the UA-Speech dataset, achieving 91.2% accuracy for ALS and 87.5% for Parkinson's Disease, outperforming MFCC features by 4.8%. Their findings highlight the effectiveness of spectrogram-based features for CNN-driven dysarthric speech recognition.

Tyagi, Dev, and Bansal (2023)[3] implemented a deep neural network (DNN) architecture, achieving a classification accuracy of 99% on dysarthric speech data. Their findings underscored the superiority of DNNs over traditional models in capturing the intricate variations present in dysarthric speech, reinforcing the potential of deep learning for highly accurate speech disorder detection. While achieving high accuracy remains a priority, ensuring computational efficiency is equally critical for real-time applications, particularly in embedded systems and mobile devices.

Addressing this challenge, Dayal et al. (2022)[4] introduced a lightweight CNN model tailored for speech signal classification. Their model delivered 95.2% classification accuracy while maintaining low computational overhead, making it highly suitable for deployment on low-power embedded systems where computational resources are limited.

Mittal et al. (2024)[5] further validated the efficacy of lightweight deep learning architectures for dysarthric speech classification. Their study employed a ResNet18-based model trained on a dataset of 2,000 dysarthric speech samples, achieving an accuracy of 96%. Their work demonstrated that carefully designed, lightweight deep learning models can offer a balance between accuracy and computational efficiency, making them viable for real-world applications in assistive communication technologies.

Additionally, Suresh, Rajan, and Thomas (2023)[6] explored the use of deep neural networks (DNNs) trained on the UA-Speech dataset, reporting a classification accuracy of 97.6%. Their results further reinforced the robustness of deep learning techniques in dysarthric speech classification, emphasizing the advantages of modern architectures over traditional machine learning approaches like Support Vector Machines (SVM) and Random Forest.

Trinh & O'Brien (2019)[7] explored CNN-based pathological speech classification using spectrogram images, achieving over 95% accuracy in distinguishing pathological from healthy speech. Their study demonstrated the robustness of CNNs for recognizing speech disorders, highlighting the effectiveness of spectrogram-based deep learning models in dysarthric speech classification.

Expanding on this approach, Musaev et al. (2019)[8] achieved 100% accuracy in Uzbek spoken digit recognition using a CNN with spectrogram-based input, demonstrating the effectiveness of image-based speech classification. Their findings support spectrogram-driven deep learning models for speech processing, extending to dysarthric speech and emotion recognition.

Akinpelu, Viriri, and Adegun (2023)[9] integrated Random Forest and Multi-layer Perceptron (MLP) within a VGGNet framework, achieving high classification accuracies of 100%, 96%, and 86.25% on the TESS, EMODB, and RAVDESS datasets, respectively. Their study highlighted the effectiveness of hybrid models for lightweight speech processing, demonstrating the feasibility of integrating deep learning with traditional classifiers for improved efficiency.

Focusing on computational efficiency, Li and Li (2023)[10] proposed a lightweight deep learning model based on depthwise separable dilated convolutions, showcasing how reducing model complexity can maintain high performance while minimizing computational costs. Their research is particularly relevant for low-power embedded applications, where resource constraints necessitate models with efficient parameter utilization and optimized inference speed.

While deep learning models achieve over 95% accuracy in dysarthric speech classification, challenges persist in balancing model size and computational efficiency. Spectrogram-based approaches, particularly Mel spectrograms, outperform MFCCs, reinforcing their effectiveness. However, lightweight architectures must be further optimized for deployment in resource-constrained environments.

III. Methodology

The proposed methodology follows a structured pipeline for dysarthria classification, involving data collection, preprocessing, spectrogram conversion, model selection, training, evaluation, and result analysis. The workflow, depicted in Fig. 1, illustrates the complete process, from raw speech input to final model selection.

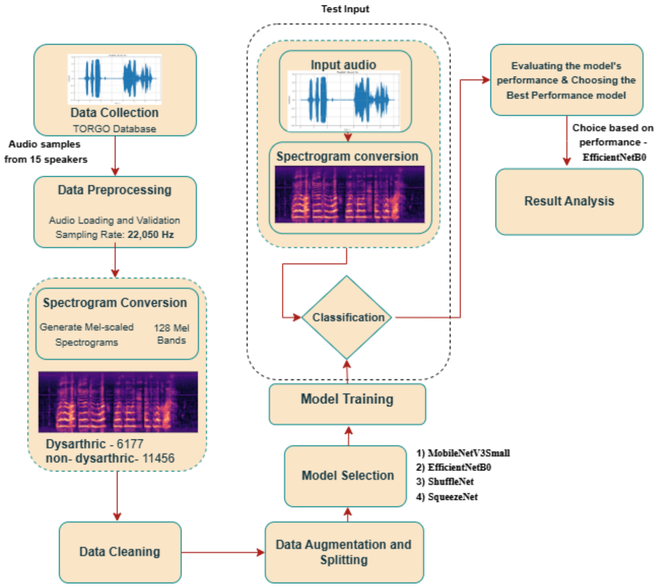

***Fig. 1**. Overview of the methodology for dysarthria classification.*

A. *Data Collection and Preprocessing*

To build a reliable dysarthria classification system, the dataset is obtained from the TORGO Database, a widely used corpus containing speech recordings from individuals with and without dysarthria. This dataset consists of speech samples from 15 speakers, capturing both male and female participants across different dysarthria severity levels. The dataset includes a total of 17,633 speech samples, comprising both dysarthric and non-dysarthric utterances. Following data acquisition, preprocessing steps are applied to standardize the audio signals. Each audio file is validated to ensure it is correctly formatted and free from corruption before being processed. The sampling rate is standardized to 22,050 Hz, ensuring consistency across all samples and aligning with standard speech processing practices. This step is crucial for maintaining uniform feature extraction during the spectrogram conversion process.

B. *Spectrogram Conversion*

Given the challenges of dysarthria classification using raw waveform data, the study employs Mel spectrogram conversion as the primary feature extraction technique. Mel spectrograms provide a time-frequency representation of speech that effectively captures phonetic variations, making them well-suited for CNN models. Each valid audio sample is transformed into a spectrogram using 128 Mel bands, which preserve critical speech features while minimizing computational overhead.

The dataset comprises 6,177 dysarthric samples and 11,456 non-dysarthric samples, leading to a noticeable class imbalance. This imbalance could introduce classification bias, causing models to favor the majority class. To mitigate this issue, data augmentation techniques are applied after the spectrogram conversion process. Techniques such as spectrogram stretching, pitch shifting, and vocal tract length perturbation are used to artificially expand the dataset, ensuring better generalization of the model across diverse dysarthric speech patterns. These augmentation strategies improve the robustness of the model, allowing it to learn variations in dysarthric speech more effectively.

The generated spectrograms are resized to 128×128 pixels, ensuring uniform input dimensions across all CNN architectures. Additionally, unnecessary background noise and artifacts are filtered out to enhance feature clarity. This structured approach to data preprocessing and augmentation significantly mitigates the issue of data scarcity, making it possible to train deep learning models effectively despite the limited availability of real-world dysarthric speech data.

C. *Selection of CNN Architecture Model*

The section on the selection of CNN architecture models describes how each of the four chosen networks—MobileNetV3Small, EfficientNetB0, ShuffleNet, and SqueezeNet—is built to achieve accurate classification of dysarthric speech while maintaining low computational requirements for deployment on devices with limited processing capabilities. Each model starts with a backbone trained on ImageNet, which provides a solid base for recognizing patterns in spectrogram images. This initial training allows the models to use previously learned visual features, reducing the need to learn everything from scratch when applied to a smaller dysarthria dataset. After the pre-trained base, additional layers are added to tailor the network for the specific task of distinguishing dysarthric from non-dysarthric speech.

MobileNetV3Small uses a method that breaks a standard convolution into two separate steps. First, it applies a filter to each channel individually, and then it combines the results from all channels. This approach cuts down the total number of mathematical operations and parameters, making the model faster and lighter qualities that are crucial for real-time use on devices like smartphones. ShuffleNet uses a design where the input channels are divided into groups that are processed separately before being mixed together. This division reduces the number of calculations while still allowing the model to combine information from different channels effectively. The model achieves a balance between accuracy and the speed required for low-power devices. SqueezeNet, on the other hand, uses a special structure that first compresses the information into a smaller set of features and then expands it through two parallel convolution operations with different kernel sizes. This design results in the fewest parameters among the four models, but it also leads to longer training times and less consistency when the data is split in different ways, suggesting that the model might have difficulty handling the variety found in spectrogram images.

EfficientNetB0 is built by adjusting the number of layers, the width of each layer, and the resolution of the input image together. It also uses specific blocks that compute a weighted average of the feature maps, which helps the network focus on the most important parts of an image. This results in the highest accuracy among the four models, though it comes with a higher number of operations and a larger model size.

D. Training and Validation

A 5-fold cross-validation strategy is implemented to ensure robust model evaluation. Cross-validation is particularly important in dysarthria classification due to the limited dataset size, as it allows the model to be trained and tested on different subsets, thereby reducing the risk of overfitting. For each fold, the dataset is split into training and validation subsets, ensuring that the model is evaluated across diverse speaker variations.

The training process involves 20 epochs with a batch size of 32. The Adam optimizer is utilized for training, providing adaptive learning rates to enhance convergence speed. The binary cross-entropy loss function is employed due to its suitability for binary classification tasks.

Several training callbacks are integrated to optimize model performance. ModelCheckpoint is used to save the best-performing model based on validation loss, ensuring that only the most accurate version of the model is retained. Additionally, EarlyStopping is implemented to halt training when validation loss stops improving, preventing unnecessary computation and reducing the likelihood of overfitting.

IV. Model implementation analysis

The comparative evaluation of MobileNetV3Small , ShuffleNet, SqueezeNet and EfficientNetB0 for dysarthria classification reveals critical insights into their performance, generalizability, and computational requirements. This analysis integrates results from 5-fold cross-validation, focusing on accuracy, precision-recall trade-offs, specificity, training dynamics, and model stability. The findings underscore the impact of architectural design on classification efficacy and practical deployment considerations.

A. Training Dynamics and Computational Efficiency

The computational efficiency of a model is crucial for real-world deployment, particularly in real time constrained environments. Table 1 presents the average epoch time, training accuracy, and early stopping behavior for each model.

Table 1: *Training Characteristics*

Model	Avg. Epoch Time (s)	Avg. Training Accuracy (%)	Early Stopping Epoch (Avg)
MobileNetV3Small	28	91.78	15
ShuffleNet	55	95.75	9
SqueezeNet	420	93.45	10
EfficientNetB0	**265**	**96.8**	**12**

EfficientNetB0 achieved the highest training accuracy (96.8%) but required 265 seconds per epoch, indicating that its training process is relatively time-consuming and better suited for settings where training is performed on dedicated hardware. ShuffleNet completed each epoch in 55 seconds while maintaining competitive accuracy (95.75%), benefiting from its channel shuffling mechanism that speeds up convergence. MobileNetV3Small was the fastest, processing an epoch in just 28 seconds, which supports its use in real-time applications, although its accuracy is lower (91.78%). SqueezeNet, despite its low parameter count, took 420 seconds per epoch, limiting its practicality for scenarios that require rapid training or frequent model updates.

B. Model Stability and Variance

A robust dysarthria classification model should deliver consistent accuracy across various data subsets to ensure it performs reliably on new, unseen speech samples. As shown in Table 2, the testing results reveal that EfficientNetB0 and MobileNetV3Small showed the most reliable performance, with accuracy standard deviations of 0.80% and 0.29%, respectively. These low values indicate that their results are stable across different folds of the dataset, making them suitable for clinical settings where speech characteristics can vary widely.

In contrast, although ShuffleNet achieved high overall accuracy, it exhibited a higher standard deviation of 1.63%. This means that its performance can vary more depending on the specific training and validation split. This fluctuation may result from the method used to mix features from grouped convolutions, which sometimes leads to less consistent feature integration. SqueezeNet displayed the highest variance (2.38%), with one-fold showing a notable drop in accuracy (89.88%) as shown in Fig 4. Such inconsistency suggests that SqueezeNet is more sensitive to variations in dataset composition or preprocessing artifacts, indicating that it may require further tuning or additional strategies to ensure stable performance in real-world scenarios.

C. Precision-Recall Trade-offs and Specificity

To further evaluate classification effectiveness, precision, recall, specificity, and F1-score were analyzed. These metrics highlight the models' ability to minimize false positives and false negatives, which is crucial in dysarthria detection. Table 2 provides a comprehensive breakdown of performance across these evaluation criteria.

Table 2: *Summary of Model Performance during testing (Mean ± Std Dev)*

Model	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	Accuracy (%)
MobileNetV3Small	91.19 ± 0.57	95.19 ± 0.34	82.92 ± 1.45	93.15 ± 0.16	90.90 ± 0.29
ShuffleNet	95.26 ± 2.39	97.28 ± 0.41	90.98 ± 4.55	96.24 ± 1.21	95.05 ± 1.63
SqueezeNet	89.78 ± 6.47	93.71 ± 4.19	93.83 ± 4.69	91.41 ± 2.37	93.15 ± 2.38
EfficientNetB0	**96.11 ± 0.92**	**96.87 ± 1.27**	**92.72 ± 1.75**	**96.47 ± 0.66**	**95.42 ± 0.80**

EfficientNetB0 and ShuffleNet demonstrated the best precision-recall trade-off, achieving high F1-scores of 96.47% and 96.24%, respectively. EfficientNetB0 's high precision (96.11%) and recall (96.87%) indicate a balanced classification approach, ensuring minimal false positives and false negatives.

ShuffleNet exhibited the highest recall (97.28%), suggesting it is highly sensitive to detecting dysarthric speech. However, its slightly lower precision (95.26%) indicates occasional false positives, which could lead to over-diagnosing non-dysarthric speech as dysarthric.

MobileNetV3Small achieved high recall (95.19%) but suffered from low precision (91.19%) and specificity (82.92%), suggesting a systemic tendency to misclassify non-dysarthric speech. SqueezeNet demonstrated competitive specificity (93.83%), but its low precision (89.78%) suggests inconsistencies in handling ambiguous spectrogram features.

D. Overall Performance and Accuracy

EfficientNetB0 emerged as the top-performing model, achieving 95.42% mean accuracy, surpassing ShuffleNet (95.05%), SqueezeNet (93.15%), and MobileNetV3Small (90.90%). Its compound scaling mechanism optimizes feature extraction across depth, width, and resolution, making it the most robust classifier for dysarthria detection. ShuffleNet demonstrated competitive accuracy despite being lightweight, reinforcing its potential for real-time applications. MobileNetV3Small 's lower accuracy highlights limitations in balancing depthwise separable convolutions, while SqueezeNet, though better than MobileNetV3Small, struggled with inconsistent generalization.

V. PERFORMANCE EVALUATION METRICS

The evaluation metrics were obtained from a comprehensive 5-fold cross-validation, with Figures 2–5 illustrating the per-fold results for MobileNetV3Small, ShuffleNet, SqueezeNet, and EfficientNetB0. The aggregated metrics—accuracy, precision, recall, specificity, and F1-score—offer a direct comparison of how each model performs in classifying dysarthric speech. Overall, the metrics indicate that while all models perform well, there are clear trade-offs. EfficientNetB0 and ShuffleNet demonstrate a strong balance between precision and recall, ensuring both high detection rates and low false positive rates. In contrast, MobileNetV3Small, although highly efficient in terms of processing speed, achieves lower overall accuracy and specificity. SqueezeNet shows greater variability in performance across different data splits, which may affect its reliability.

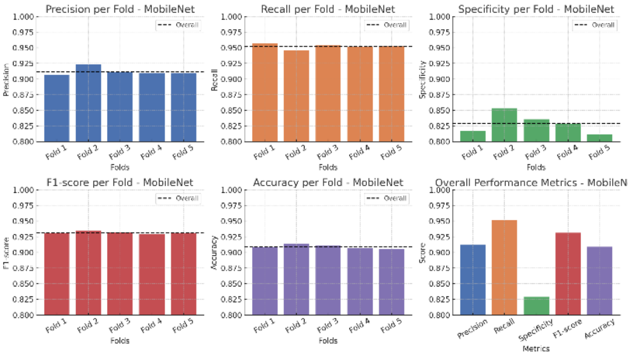

Fig 2. *per-fold and overall performance metrics for MobileNet*

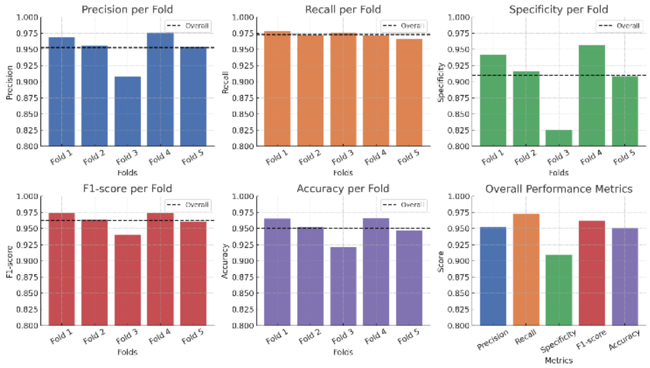

Fig 3. *per-fold and overall performance metrics for ShuffleNet*

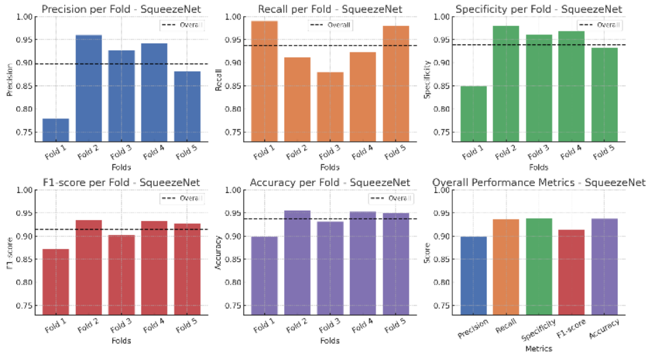

Fig 4. *per-fold and overall performance metrics for SqueezeNet*

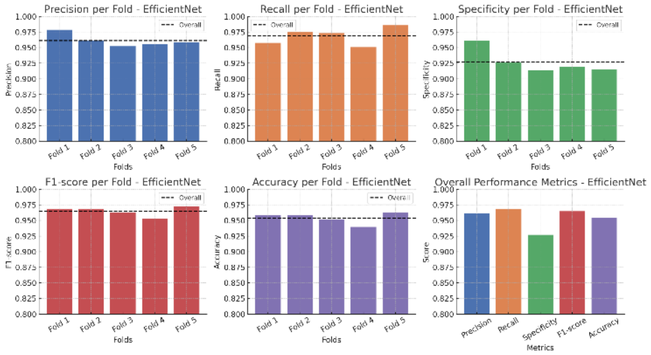

Fig 5. *per-fold and overall performance metrics for EfficientNet*

VI. CONCLUSION

This study evaluated four lightweight deep learning models—MobileNetV3Small, ShuffleNet, SqueezeNet, and EfficientNetB0 for binary dysarthric speech classification using spectrogram images and 5-fold cross-validation. Each model was assessed in terms of accuracy, computational efficiency, and total parameters, providing insights into their suitability for real-time applications. MobileNetV3Small, optimized for speed, achieved 90.90% accuracy with only 1.01 million total parameters (73,985 trainable). While it provides rapid inference for real-time applications, the reduced model complexity leads to a slight drop in classification specificity, making it better suited for scenarios requiring low-latency processing rather than high precision. ShuffleNet, designed for high efficiency, attained a comparable accuracy (95.05%) while maintaining a lower parameter count (1.81 million total, 1.79 million trainable). By using group convolutions and channel shuffling, it minimizes computation without sacrificing accuracy, making it ideal for embedded systems and edge devices where power efficiency is a concern. SqueezeNet, despite its ultra-lightweight architecture (723,009 parameters, all trainable), exhibited variable performance (93.15% accuracy) and slower convergence. The fire modules helped reduce parameters, but the compact design sometimes limited its feature extraction capabilities. As a result, SqueezeNet is best suited for highly constrained environments, such as IoT devices, where minimal storage and computational power are available. EfficientNetB0 achieved the highest accuracy (95.42%) and the lowest loss (0.1321) due to its compound scaling strategy, which balances depth, width, and resolution. However, its high computational cost (7.2 million parameters, with 3.15 million trainable) makes it less feasible for real-time or mobile applications, favoring cloud-based deployment instead. In summary, our work comprehensively evaluates lightweight CNNs for dysarthric speech classification, balancing accuracy, efficiency, and deployment feasibility. ShuffleNet and EfficientNetB0 offer high performance, while MobileNetV3Small and SqueezeNet suit resource-constrained scenarios. Limitations include a trade-off between complexity and precision, and potential performance variability in ultra-light models.

A. Future Works

Future research should focus on enhancing dataset diversity to improve generalization across dysarthric severity levels and speaker variations. Advanced data augmentation techniques (e.g., spectrogram stretching, pitch shifting, vocal tract length perturbation) can further improve model robustness. Expanding classification beyond binary detection to severity-level classification would enhance clinical assessments.

Exploring hybrid lightweight CNNs, attention-based models (e.g., Squeeze-and-Excitation networks), and alternative feature extraction methods like MFCCs or learned embeddings could improve feature representation. Optimizing models through quantization, pruning, and hardware-aware design is essential for real-time deployment. Integrating transfer learning (e.g., Whisper encoder) could enhance performance while reducing data and training time.

Future studies should incorporate longitudinal evaluations, multilingual datasets, and domain adaptation to ensure robustness across diverse linguistic contexts. Expanding evaluation metrics beyond accuracy and loss (e.g., precision, recall, F1-score, AUC) would provide a more comprehensive assessment. Explainable AI (XAI) techniques can improve model transparency, while addressing class imbalances through data augmentation and class-weighted loss functions would enhance fairness.

These advancements will refine DSR models, making them more accurate, efficient, and accessible for clinical applications and assistive communication technologies.

References

[1] A. A. Joshy and R. Rajan, "Automated Dysarthria Severity Classification Using Deep Learning Frameworks," 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 2021, pp. 116-120, doi: 10.23919/Eusipco47968.2020.9287741.

[2] B. Suhas et al., "Speech task based automatic classification of ALS and Parkinson’s Disease and their severity using log Mel spectrograms," 2020 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 2020, pp. 1-5, doi: 10.1109/SPCOM50965.2020.9179503.

[3] V. Tyagi, A. Dev, and P. Bansal, "Analysis and Classification of Dysarthric Speech," 2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Delhi, India, 2023, pp. 1-6, doi: 10.1109/O-COCOSDA60357.2023.10482956.

[4] A. Dayal, S. R. Yeduri, B. H. Koduru, R. K. Jaiswal, J. Soumya, M. B. Srinivas, O. J. Pandey, and L. R. Cenkeramaddi, "Lightweight deep convolutional neural network for background sound classification in speech signals," J. Acoust. Soc. Am., vol. 151, no. 4, pp. 2773–2786, 2022, doi: 10.1121/10.0010257.

[5] K. Mittal, K. S. Gill, K. Rajput, and V. Singh, "Enhancing the Diagnosis of Speech Disorders: An In-Depth Investigation into Dysarthria Classification Using the ResNet18 Model," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-5, doi: 10.1109/ICITEICS61368.2024.10625627.

[6] M. Suresh, R. Rajan, and J. Thomas, "Dysarthria Speech Disorder Classification Using Traditional and Deep Learning Models," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023, pp. 01-06, doi: 10.1109/ICEEICT56924.2023.10157285.

[7] Trinh, N., O'Brien, D. (2019). Pathological speech classificaiton using a convolutional neural network. IMVIP 2019: Irish Machine Vision & Image Processing, Technological University Dublin, Dublin, Ireland, August 28-30. doi:10.21427/9dnc-n002

[8] M. Musaev, I. Khujayorov, and M. Ochilov, "Image Approach to Speech Recognition on CNN," in Proc. 2019 3rd International Symposium on Computer Science and Intelligent Control (ISCSIC 2019), Association for Computing Machinery, New York, NY, USA, Article 57, pp. 1–6, 2020, doi: 10.1145/3386164.3389100.

[9] S. Akinpelu, S. Viriri, and A. Adegun, "Lightweight Deep Learning Framework for Speech Emotion Recognition," IEEE Access, vol. 11, pp. 77086-77098, 2023, doi: 10.1109/A8CCESS.2023.3297269.

[10] Z. Li and W. Li, "MOSLight: A Lightweight Data-Efficient System for Non-Intrusive Speech Quality Assessment," in Proc. Interspeech 2023, pp. 5386-5390, 2023, doi: 10.21437/Interspeech.2023-263.

[11] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Dysarthric Speech Recognition Based on Deep Metric Learning," in Proc. Interspeech 2020, pp. 4796-4800, 2020, doi: 10.21437/Interspeech.2020-2267.

[12] D. Mulfari, A. Celesti, and M. Villari, "Exploring AI-based Speaker Dependent Methods in Dysarthric Speech Recognition," 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Taormina, Italy, 2022, pp. 958-964, doi: 10.1109/CCGrid54584.2022.00117.

[13] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140.

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6848-6856, doi: 10.1109/CVPR.2018.00716.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5MB model size," arXiv preprint arXiv:1602.07360, 2016.

[16] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105-6114.