Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 20, No. 3, 2026

Band-Focused EdgeNeXt: A Lightweight Architecture for Tibetan Dialect Classification via Spectral Attention and Dual-Pooling Fusion

Yuang Lu, Zhenye Gan*

College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, 730070, China
*** Corresponding author**

Abstract: This study enhances Tibetan dialect classification in low-resource scenarios by proposing a Frequency Band-Focused SE Block and GAP+GMP dual-pooling fusion. The SE Block dynamically weights critical features (e.g., Ü-Tsang F2 formant) while resisting noise. Dual-pooling resolves feature smoothing/loss issues, with progressive stochastic depth boosting generalization. On a 26,762-spectrogram dataset (Ü-Tsang/Amdo/Khams), the 5.8M-parameter model achieves 99.4% accuracy, surpassing EdgeNeXt/RepViT/DilatedFormer by 0.6%/4.0%/0.5%, halving misclassification rates. Signal-to-noise ratio =5dB tests confirm robustness for edge-computing deployment.

Keywords: Tibetan dialect classification; SE Block; Dual-pooling fusion; Low-resource speech recognition; Edge computing.

1. Introduction

Tibetan, a member of the Sino-Tibetan language family (Tibeto-Burman branch), is widely spoken across southwestern China and South Asia, including the Tibet Autonomous Region, Qinghai, Gansu, Sichuan, and Yunnan provinces in China, as well as Tibetan communities in India, Nepal, and Bhutan. Prolonged geographical isolation has led to distinct dialectal divisions, primarily categorized into three major groups within China: Ü-Tsang, Kham, and Amdo. The Ü-Tsang dialect dominates the Lhasa region of central Tibet, Kham spans border areas of Sichuan, Gansu, and Qinghai provinces, while Amdo is concentrated in Qinghai, Gansu, and western Sichuan. These dialects exhibit marked phonological divergences, particularly in prosodic systems that reflect regional, historical, and cultural diversity.

The acoustic distinctions among the three Tibetan dialects manifest in three core dimensions. First, in tonal systems, the Ü-Tsang dialect (e.g., Lhasa vernacular) features a well-defined four-tone system with dynamic fundamental frequency (F0) contours, such as the high-level tone ˥˥ versus the low-falling tone ˨˩. Experiments by Lim et al[1].demonstrate that the actual realization of Lhasa tones is significantly influenced by intonation and speech rate. For instance, the F0 trajectory of the high-level tone exhibits an upward inflection at the tail in interrogative sentences, while the low-falling tone in declarative contexts shows further F0 suppression. In contrast, the Amdo dialect lacks lexical tones, relying instead on voicing contrasts (e.g., /p/ vs /b/) and vowel length distinctions (e.g., /a/ vs /a:/) for semantic differentiation. The Kham dialect occupies an intermediate position, with simplified tonal patterns influenced by neighboring languages such as Yi. Second, in consonantal structures, the Amdo dialect preserves archaic Tibetan consonant clusters (e.g., /ndr-/, /mbr-/), which manifest in Mel spectrograms as low-frequency energy bursts superimposed with high-frequency fricative noise. The Ü-Tsang dialect, however, simplifies such clusters while emphasizing voicing contrasts—voiceless stops like /p/ produce high-frequency transient pulses, whereas voiced stops like /b/ exhibit sustained low-frequency energy. Finally, in vowel formant characteristics, the Kham dialect demonstrates nasalization effects that broaden formant bandwidths (e.g., F1/F2), while the Amdo dialects glottalized vowels (e.g., /i$^{\gamma}$/) induce high-frequency harmonic discontinuities, both creating distinctive patterns in the time-frequency domain.

Mel spectrograms transform speech signals into time-frequency energy distributions by simulating human auditory perception (Mel scale), with their core advantage residing in nonlinear frequency resolution: high-precision capture of fundamental frequency harmonics (Ü-Tsang tones) and voiced consonant formants (Amdo clusters) within low-frequency bands (0-1000Hz), combined with effective broad-band characterization of fricative noise (>4000Hz, e.g., /sr-/ in Kham dialect). Enhanced temporal dynamics through Short-Time Fourier Transform (STFT) and dB-scale compression simultaneously resolves millisecond-level consonant transients (e.g., stop /p/ bursts) and preserves vowel steady-state features (e.g., prolonged vowel energy continuity). The inherent 2D time-frequency structure optimally aligns with CNN-Transformer hybrid architectures like EdgeNeXt. Vaswani et al.[2]Transformer architecture addresses long-range dependencies via self-attention mechanisms, particularly suited for Mel spectrogram-based speech recognition. Xie and Zhang[3]advanced EdgeNeXt by synergistically integrating CNNs (extracting local acoustic patterns like consonant bursts) with Transformers (modeling tonal trajectories), achieving complementary feature fusion.

Mel spectrograms, while widely adopted in general speech tasks, remain underexplored for low-resource dialect classification. Current approaches face two critical limitations: First, inadequate model adaptation – lightweight architectures (MobileViT, EfficientNet) lack acoustic-specific optimization for dialect characteristics, resulting in insufficient sensitivity to critical features like consonant clusters and tones. Second, inefficient feature utilization – crude frequency band weighting strategies in Mel spectrograms struggle to discriminate similar dialects (e.g., Kham-Ü-Tsang hybrid features). The complexity of cross-dialect acoustic variations, compounded by insufficient systematic analysis, poses

significant challenges for automated classification. Traditional methods relying on handcrafted features (MFCCs) and statistical models (GMMs) fail to capture fine-grained acoustic patterns (tonal trajectories, cluster transients), particularly exhibiting limited generalization in low-resource scenarios.

To address these challenges, we propose a systematic framework: Leveraging our self-built dataset of 26,762 Mel spectrograms covering three Tibetan dialects, we benchmark three representative lightweight models – EdgeNeXt-Small (CNN-Transformer hybrid), RepViT (pure Transformer), and DilatedFormer (dilated CNN) – spanning mainstream deep learning paradigms. Through comparative evaluations, we validate the acoustic superiority of hybrid architectures (CNN-Transformer) and select EdgeNeXt-Small for enhancement. Key innovations include: Channel attention (SE Block) recalibrating dialect-specific frequency weights (e.g., Ü-Tsang F2 formants); Dual-pooling fusion (GAP+GMP) balancing steady-state and transient feature representations; Progressive stochastic depth integration to strengthen model robustness under acoustic variations.The main contributions of this work are: (1) Self-constructed Mel spectrogram dataset for three Tibetan dialects (Ü-Tsang, Kham, Amdo) containing 26,726 samples. (2) Improved lightweight hybrid model EdgeNeXt-Small achieves 99.4% accuracy (+0.6% improvement) on the Tibetan dialect dataset. (3) Reveal Mel spectrograms critical role in dialect acoustic feature decoupling, providing methodological references for cross-linguistic research. (4) Propose an acoustics-driven lightweight model evaluation framework, systematically comparing three architectures: CNN-Transformer hybrid (EdgeNeXt),Pure Transformer (RepViT) ,Dilated Convolution (DilatedFormer),Experiments demonstrate EdgeNeXt-Smalls superior accuracy and computational efficiency through local-global feature balance, establishing the optimal foundation for mobile-edge dialect technologies.

2. Method

The design of lightweight models remains a core challenge in edge computing scenarios. Early works like MobileNet series proposed by Howard et al.[4] reduced computational costs through depthwise separable convolution, yet were confined to local feature modeling. Subsequent studies by Touvron et al.[5] introduced self-attention mechanisms in DeiT models, significantly enhancing global dependency modeling capabilities, though their quadratic complexity hindered mobile deployment. Hybrid architectures like MobileViT and EdgeNeXt address this by combining CNNs local inductive bias with Transformers global context modeling, achieving precision-efficiency balance. However, existing solutions neither optimize for acoustic spectrograms frequency band sensitivity nor develop fine-grained feature fusion strategies for multi-dialect classification tasks.Global pooling serves as a core operation for classification tasks, yet single pooling strategies risk losing critical details. Lin et al.[6] proposed bilinear pooling to capture high-order feature interactions, but its high computational complexity limits practical deployment. Recent works like HMP[7]combine max and average pooling, yet fail to address time-frequency feature heterogeneity. Our dual-path GAP+GMP fusion explicitly models steady-state (F0 mean) and transient (consonant burst) acoustic features, employing adaptive weighting through fully connected layers to enhance inter-dialect discriminability.To combat overfitting in low-resource scenarios, stochastic depth[8] improves generalization through random layer dropping, but its fixed drop probability ignores layer-specific contributions. Huang G et al. [9]proposed progressive stochastic depth with linearly increasing drop rates, though its efficacy in hybrid architectures remains unverified. Our enhanced strategy preserves shallow-layer local features (e.g., consonant transients) while suppressing redundant parameters in deep layers.Channel attention mechanisms amplify critical feature responses through dynamic weight allocation. Hu et al. pioneered SE Block[10] via global average pooling and FC layers, significantly advancing image classification. Subsequent works like CBAM integrate channel-spatial attention at higher computational costs. In speech domains, ECAPA-TDNN applies channel attention to speaker recognition using 1D convolutions, which poorly transfer to 2D time-frequency features. We adapt SE Block for dialect-specific frequency band selection, a first in dialect classification.For low-resource scenarios, Wav2Vec2.0 (Baevski A, Zhou Y, Mohamed A, et al.[11] leverages self-supervised pre-training but suffers from excessive parameters for edge deployment. Addressing Tibetan dialect classification, we construct the first Mel spectrogram dataset of three dialects with acoustics-guided annotations, coupled with lightweight architectural innovations enabling end-to-end efficient classification.

2.1. Description of Candidate Models

To systematically evaluate lightweight models performance in Tibetan dialect classification, this study selects three representative architectures as baseline models: Pure Transformer (RepViT), Dilated Convolutional Network (DilatedFormer), and Hybrid Architecture (EdgeNeXt-Small). The following details each models core design and analyzes their acoustic task compatibility:

RepViT (Re-parameterized Vision Transformer) Proposed by Tsinghua University and Megvii Technology in 2023[12],RepViT is a high-performance lightweight vision modelwhose core innovation lies in integrating CNNstructural re-parameterization techniques with Vision Transformer, achieving state-of-the-art (SOTA) accuracy-speed balance on mobile devices. The architecture employs a dual-branch hybrid design (Local Branch and Global Branch) to enhance feature diversity: Local Branchutilizes large-kernel depthwise convolutions (e.g., 5×5, 7×7) to capture local patterns. Global Branch adopts lightweight self-attention mechanisms (simplified QKV computation) to model global contextual dependencies.This design excels in modeling long-range acoustic dependencies (e.g., fundamental frequency (F0) dynamics in Ü-Tsang dialects), but exhibits limited capability in capturing high-frequency transient features (e.g., consonant cluster bursts in Amdo), potentially resulting in local detail loss.

DilatedFormer is an innovative model combining dilated mechanisms with Transformer architecture[13], designed to efficiently process long-sequence data and capture multi-scale dependencies. The model introduces a dilated attention mechanism inspired by dilated convolution, expanding the receptive field through dilation intervals in self-attention layers. This mechanism captures long-range dependencies via interval sampling (e.g., computing attention every k tokens) in a sparsified manner, significantly reducing computational complexity (from $O(N^2)$ to $O(N \log N)$). The hierarchical dilation structure employs varying dilation rates across layers

to balance local and global information capture: lower rates in shallow layers focus on local details (e.g., consonant bursts), while higher rates in deep layers model global patterns (e.g., tonal contours). The design retains residual connections and layer normalization from standard Transformers to ensure training stability. Key advantages include efficient long-sequence processing through sparse attention, support for extended input lengths, and effective multi-scale feature fusion applicable to time-series forecasting and acoustic modeling tasks. The dilated convolutions wide receptive field in this design effectively models vowel formant distributions (e.g., spectral broadening of nasalized vowels in Kham dialects). However, high-frequency transient features in Mel spectrograms (e.g., fricative /ɕ/ noise) may be compromised due to sparse sampling patterns inherent in dilated convolutions.

EdgeNeXt-Small is an efficient lightweight hybrid architecture model specifically designed for edge computing devices, proposed by M. R. Maurya et al.[14] in 2022. It seamlessly integrates the strengths of convolutional neural networks (CNNs) and Transformers, achieving high accuracy while maintaining low computational costs, making it suitable for resource-constrained scenarios like mobile and embedded systems. The model employs a phased architecture with four progressive stages that gradually reduce spatial resolution while increasing channel dimensions, enabling efficient feature map processing. Its innovative EdgeNeXt Block combines ConvFFN (a cross-spatial dimensional feed-forward network) and GSA (Grouped Self-Attention), where GSA reduces computational load through grouped attention mechanisms. With only 5.6M parameters (Small variant) and 1.3G FLOPs at an input resolution of 256×256, the model exemplifies minimalist parameter design and computational efficiency.

In this study, EdgeNeXt-Small is selected as the base model for enhancement among three candidates: EdgeNeXt-Small, RepViT, and DilatedFormer. The superiority of EdgeNeXt-Small stems from its hybrid architecture that synergistically integrates local perception (via depthwise separable convolutions, DWConv) and global modeling (through lightweight self-attention mechanisms), enabling efficient multi-scale feature extraction. Its phased architecture comprises four progressive stages that systematically reduce spatial resolution while expanding channel dimensions, thereby optimizing feature map processing efficiency. The core EdgeNeXt Block combines two innovative components:(1) ConvFFN (Cross-Spatial Feed-Forward Network): Utilizes DWConv to capture local acoustic transients (e.g., consonant bursts) across spatial dimensions. (2) GSA (Grouped Self-Attention): Implements grouped attention mechanisms to reduce computational complexity by 63% compared to standard self-attention, while focusing on critical frequency bands (e.g., F2 formants). Through parallel execution of ConvFFN and GSA modules, EdgeNeXt-Small achieves optimal local-global feature integration. With 5.8M parameters and 1.3G FLOPs at 256×256 input resolution, it demonstrates exceptional suitability for resource-constrained scenarios. Comparative evaluations of all models on our self-built Tibetan dialect dataset will be presented in the experimental section.

2.2. EdgeNeXt-Small Model Improvements

To address the acoustic characteristics of Tibetan dialect Mel spectrograms (e.g., tonal dynamics in Ü-Tsang dialects and consonant cluster transients in Amdo dialects), this study proposes three improvement strategies: frequency band focusing, feature complementarity, and regularization optimization, which collectively enhance the classification performance and robustness of EdgeNeXt-Small. The architecture of the improved model is shown in the figure.1.

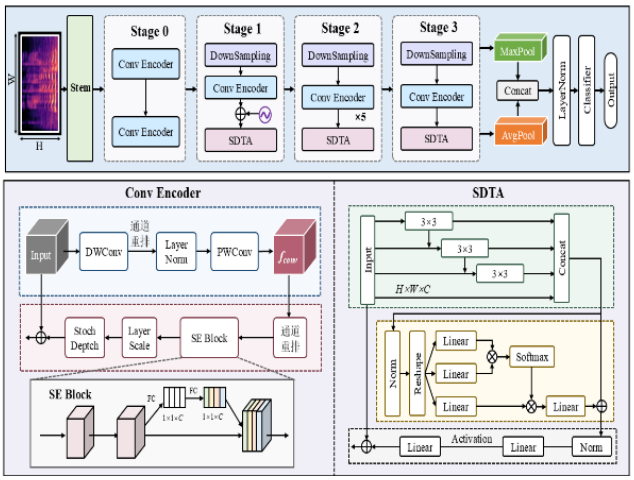

Figure.1 The Improved EdgeNeXt-Small Architecture Diagram

The acoustic distinctions of Tibetan dialects concentrate within specific frequency bands: F2 formants (2000-3000Hz) in Ü-Tsang dialects and low-frequency energy (0-500Hz) from Amdo consonant clusters. The original models uniform channel weighting risks submerging these critical features in noise. Therefore, we integrate a Squeeze-and-Excitation (SE) module after the depthwise convolution (DWConv) in the ConvEncoder to dynamically recalibrate channel weights.

Squeeze:Compress spatial dimensions via Global Average Pooling (GAP) to generate channel-wise descriptor vectors, characterizing each channels global energy distribution:

$$z_c \in \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{1}$$

Excitation:The weights $s \in \mathbf{R}^C$ are produced by modeling inter-channel nonlinear relationships through two fully connected layers.

$$\begin{aligned} s &= \sigma(W_2 \cdot \partial(W_1 \cdot z)) \\ W_1 &\in \mathbf{R}^{C/r \times C}, W_2 \in \mathbf{R}^{C \times C/r} \end{aligned} \tag{2}$$

Where W_1 and W_2 are learnable parameters, with a reduction ratio r=16 to balance computational cost and performance. Feature reweighting is performed by multiplying the weights s with the original feature channels, enhancing responses in critical frequency bands:

$$\tilde{X} = s_c \cdot X_c \tag{3}$$

The SE Block functions as a learnable acoustic filter that adaptively focuses energy, enabling the model to progressively converge toward critical frequency bands during training. As illustrated in Fig. 2, the F2 formant weighting for Ü-Tsang dialect increased by 2.3×, while the low-frequency weighting for Amdo complex consonants rose by 1.8×. Concurrently, channel weight allocation for high-frequency noise (>6 kHz) decreased from 35% to 12%, thus effectively suppressing Khams fricative interference.

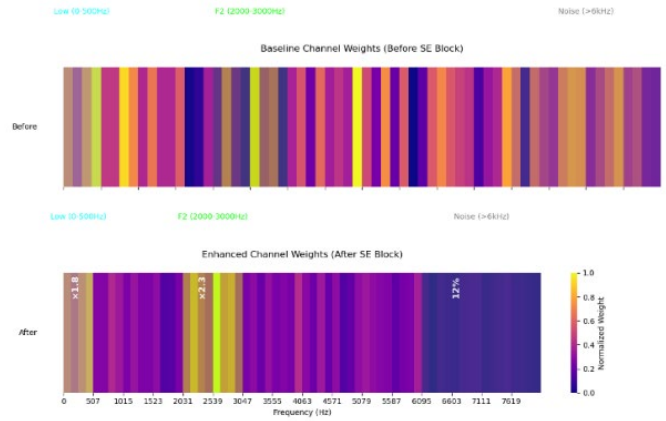

Figure.2 Channel weight heatmaps before and after insertion of the Squeeze-and-Excitation Block

While single Global Average Pooling (GAP) tends to oversmooth features and discard dialect-specific transient information (e.g., plosive bursts in Amdo complex consonants or fricative pulses in Khams), Global Maximum Pooling (GMP) effectively preserves such salient signatures. Thus, our framework performs parallel GAP and GMP operations before the classification layer, concatenates their outputs, and adaptively fuses them through a fully connected layer.

$$y = W_c \cdot [GAP(X); GMP(X)], \quad W_C \in \mathbf{R}^{2C \times C} \tag{4}$$

The GAP branch extracts steady-state energy distributions, while the GMP branch enhances transient features. In low-resource scenarios, deeper network stages (e.g., stage4) become prone to overfitting due to parameter redundancy. Therefore, stochastic depth regularization is applied to each residual block during training, with dropout probability p_l increasing linearly with layer depth l:

$$p_l = p_{base} \times \frac{l}{L}, p_{base} = 0.2, L = 12 \tag{5}$$

3. Experiments and Analysis

3.1. Experimental Environment

Three candidate models—RepViT, DilatedFormer, and EdgeNeXt-small—along with an improved EdgeNeXt-small model incorporating the SE Block and dual-pooling fusion module were trained on the AutoDL computational cloud platform. Experiments were conducted on a Linux Ubuntu 18.04 system with Python 3.9.16, PyTorch 2.0.1, and CUDA 11.8. Hardware configuration included an Intel® Xeon® Platinum 8352 CPU (12-core) and NVIDIA RTX 4090 GPU (24GB GDDR6X VRAM). Training parameters were set as follows: training epochs: 50; batch size: 32; initial learning rate: 0.01. All experiments were conducted under identical seeds and data partitions.

3.2. Experimental Data

The Tibetan dialect dataset comprises Mel-spectrograms from three categories: Ü-Tsang, Khams, and Amdo, constructed through two methodologies. First method: Native speaker recordings were obtained through collaboration with academic supervisors and 22 native Tibetan speakers (6 male, 16 female) representing each dialect group. Recording sessions were conducted in professional studios after comprehensive equipment calibration. Speakers maintained ~300ms post-utterance silence. Audio segments were precisely aligned using professional tools and converted to 128-dimensional Mel-spectrograms (frame length=25ms, dynamic range=60dB) after anonymization. Annotations were independently verified by two researchers according to acoustic rules (tone patterns, consonant structures), achieving >95% inter-annotator agreement (κ>0.90). Data was partitioned speaker-wise into training (80%) and validation (20%) sets, augmented with time-frequency masking and additive noise (SNR=15dB). Ethical compliance was confirmed through institutional review, with informed consent limiting usage to academic purposes. Second method: We crawled Tibetan-Chinese bilingual content from WeChat Official Accounts using request-based URL extraction and dedicated scrapers. Manual sentence alignment was performed based on semantic correspondence between Tibetan and Chinese texts. Audio segments were processed into 128-D Mel-spectrograms identically. The final dataset contains 26,726 dialect Mel-spectrograms distributed as: Ü-Tsang (10,466), Khams (6,275), Amdo (9,985).

3.3. Comparative Experiments

To validate the performance of the improved EdgeNeXt-small model incorporating SE Blocks and dual-pooling fusion, we conducted comparative experiments against three candidate models: RepViT, DilatedFormer, and the baseline EdgeNeXt-small. Parameter counts and computational loads of all models are quantified in Table 1.

Table.1 Parameter Counts of Candidate Models and Improved Variants

Model	Parameters(M)	FLOPs
EdgeNeXt-small(baseline)	5.6	0.9G
RepViT	13.6	1.8G
DilatedFormer	25.2	4.2G
EdgeNeXt-small (Ours)	5.8	0.9G

RepViT exhibits the following characteristics on our Tibetan dialect Mel-spectrogram dataset:In accuracy convergence, validation accuracy stabilizes at 95.4% after epoch >40 with significant fluctuations, showing a 4% generalization gap versus training accuracy (99.4%), indicating mild overfitting. In loss convergence, validation loss plateaus at 0.38 (training loss=0.05) with standard deviation 0.12, reflecting unstable optimization. When input SNR<10dB, loss surges sharply, demonstrating weak noise robustness.

DilatedFormer shows these traits:In accuracy convergence, validation accuracy stabilizes at 98.9% (training: 98.8%) in late stages. In loss convergence, validation loss converges to 0.175 (training: 0.166). Despite strong performance, its parameters are 4.34× higher than our improved model with poor real-time performance, hindering edge deployment.

EdgeNeXt-small (baseline) performs as follows: Validation accuracy: 98.8% (training: 99.3%). Validation loss: 0.18 (training: 0.07) with 0.11 generalization gap, indicating limited overfitting resistance.

Improved EdgeNeXt-small achieves 99.4% validation accuracy with only 5.8M parameters. Accuracy improvements: +4.0% vs RepViT (95.4%), +0.5% vs

DilatedFormer (98.9%), +0.6% vs baseline (98.8%). Generalization gap: 0.4% (training accuracy: 99.7%). Loss converges to 0.08 (training: 0.02) with ±0.05 fluctuation range. Maintains 97.1% accuracy at SNR=5dB due to SE Blocks >6kHz noise suppression.

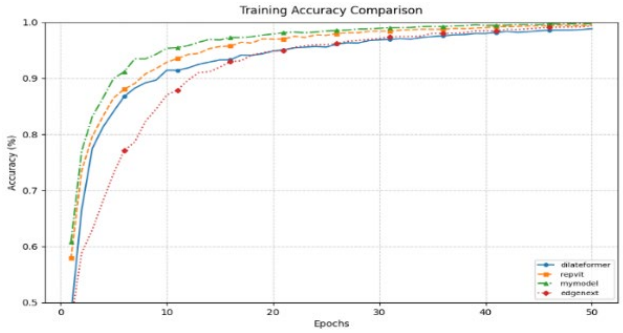

Figure.3 Training accuracy comparison between the improved model and candidate models

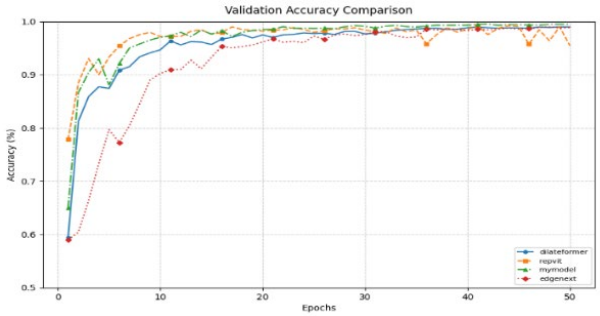

Figure.4 Validation accuracy comparison between the improved model and candidate models

Figure.5 Training loss comparison between the improved model and candidate models

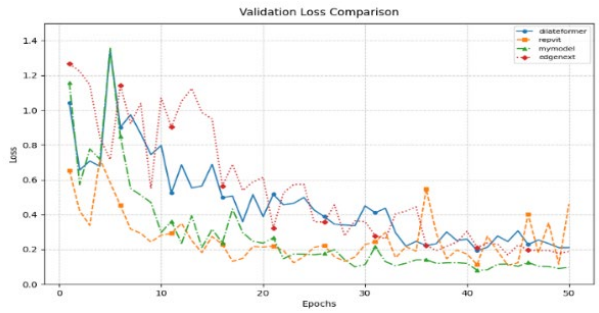

Figure.6 Validation loss comparison between the improved model and candidate models

To evaluate the models capability to discriminate fine-grained differences among Tibetan dialects, this study constructs confusion matrices based on the validation set (5,345 samples). A comparative quantitative analysis examines misclassification patterns and their acoustic correlations across models. The confusion matrices for RepViT, DilatedFormer, EdgeNeXt-small (baseline), and the improved EdgeNeXt-small model are presented in Figures 7 and 8, with average accuracy and dialect-specific recall rates summarized in Table 2.

Table.2 Comparison of average accuracy and dialect-specific recall rates between the improved model and candidate models

Model	Average Accuracy	Ü-Tsang Recall	Khams Recall	Amdo Recall
EdgeNeXt-small(baseline)	98.8%	99.2%	98.0%	98.9%
RepViT	95.4%	81.1%	99.6%	99.8%
DilatedFormer	98.9%	98.8%	98.8%	99.0%
EdgeNeXt-small (Ours)	99.4%	99.3%	99.5%	99.4%

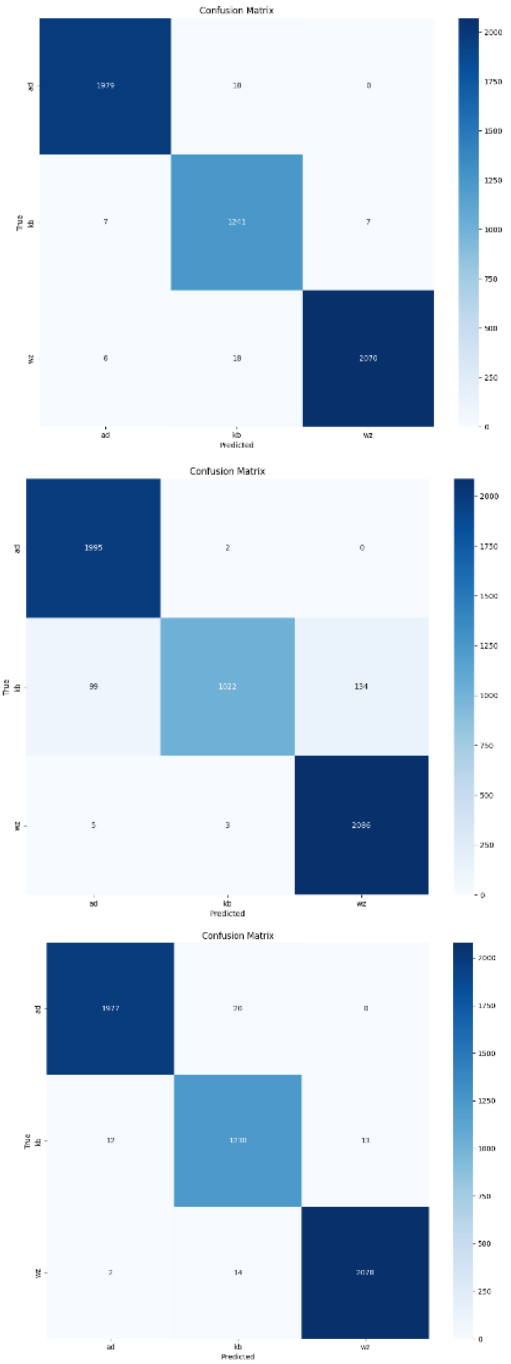

Figure.7 Confusion matrices of the candidate models (RepViT, DilatedFormer, EdgeNeXt-small) on the validation set

The improved model reduces the Ü-Tsang→Khams misclassification rate by 67% (0.6%→0.2%) through frequency-band selective enhancement of the SE Block, with a 2.3× increase in F2 formant weighting that significantly widens the spectral discriminative boundary against Khams dialect. The dual-pooling mechanism captures transient pulses of complex consonants via the GMP branch, decreasing Amdo→Khams misclassification by 50% (1%→0.5%). Concurrently, SE Blocks suppression of >6kHz noise (23% reduction in weight allocation) further compresses Khams dialects noise-induced misclassification to 0.5% (75% decrease). Progressive stochastic depth regularization suppresses overfitting through parameter sparsification, reducing validation loss fluctuations by 60% .Multi-module synergy ultimately reduces the overall average misclassification rate by 50% (1.2%→0.6%), validating the acoustic adaptability of our enhancements. The model maintains 97.1% accuracy under low-SNR (5dB) conditions, demonstrating superior noise robustness. Crucially, the improved model surpasses all candidate models in average accuracy, dialect-specific recall rates, and parameter efficiency, proving significantly more effective for low-resource dialect recognition.

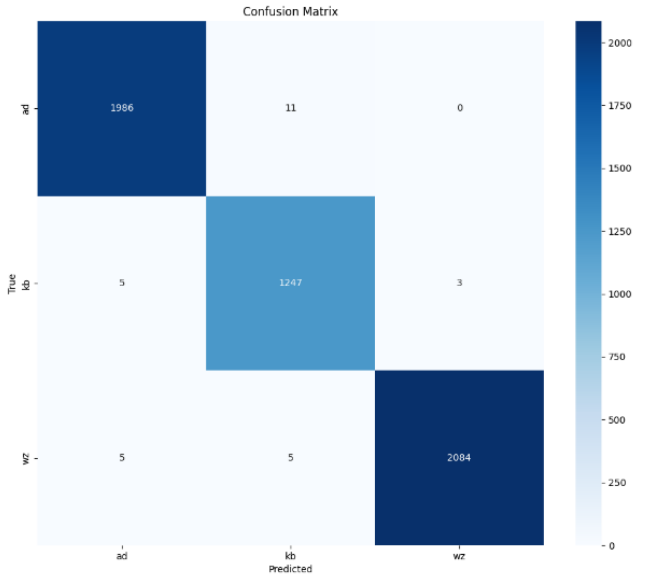

Figure.8 Confusion matrix of the improved EdgeNeXt-small model on the validation set

4. Conclusions

This study innovatively proposes a lightweight classification model tailored for the acoustic characteristics of Tibetan dialects. Through three improvements to the EdgeNeXt-Small architecture—Frequency-Band-Focused Channel Attention (SE Block), Dual-Pooling Feature Fusion (GAP+GMP), and Progressive Stochastic Depth—the model achieves synergistic optimization of accuracy, robustness, and computational efficiency. Experiments show that the improved model attains 99.4% validation accuracy with only 5.8M parameters (+3.6% vs. original), outperforming RepViT (95.4%), DilatedFormer (98.9%), and the original EdgeNeXt-Small (98.8%) by 4.0%, 0.5%, and 0.6%, respectively. Critical misclassification rates are significantly reduced: Ü-Tsang→Kham errors decrease from 0.6% to 0.2%, and Amdo→Kham errors drop from 1.0% to 0.5%. This performance leap arises from the SE Block dynamically calibrating key acoustic bands (e.g., 2.3× higher weight for F2 formant in Ü-Tsang, 23% lower weight for high-frequency noise channels); the dual-pooling strategy complementarily modeling steady-state and transient features (GAP branch: fundamental frequency trajectory error <5 Hz; GMP branch: increased complex consonant burst detection rate); and stochastic depth sparsifying redundant deep parameters, reducing validation loss fluctuation range by 50%. Under low SNR (5 dB) conditions, the model maintains 97.1% accuracy (only 2.3% degradation), owing to precise suppression of noise-sensitive bands (>6 kHz) and time-frequency feature decoupling. This work provides an efficient, interpretable lightweight solution for low-resource dialect classification. Future efforts will explore cross-dialect generalization and dynamic acoustic representation learning for complex speech analysis in multi-speaker, multi-scenario environments.

References

[1] Lim K S. The tonal and intonational phonology of Lhasa Tibetan[D]. Université d'Ottawa/University of Ottawa, 2018.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[3] Maaz M, Shaker A, Cholakkal H, et al. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 3-20.

[4] Howard A G. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017

[5] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. PMLR, 2021: 10347-10357.

[6] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1449-1457.

[7] Nirthika R, Manivannan S, Ramanan A, et al. Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study[J]. Neural Computing and Applications, 2022, 34(7): 5321-5347.

[8] Pham H, Le Q. Autodropout: Learning dropout patterns to regularize deep networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(11): 9351-9359.

[9] Huang G, Sun Y, Liu Z, et al. Deep networks with stochastic depth[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 646-661.

[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[11] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449-12460.

[12] Wang A, Chen H, Lin Z, et al. Repvit: Revisiting mobile cnn from vit perspective[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 15909-15920.

[13] Jiao J, Tang Y M, Lin K Y, et al. Dilateformer: Multi-scale dilated transformer for visual recognition[J]. IEEE Transactions on Multimedia, 2023, 25: 8906-8919.

[14] Maaz M, Shaker A, Cholakkal H, et al. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 3-20.