Journal of Modern Researches on Electronic Defense Systems

Receive: 4 January 2026 Accept:12 June 2026

Automatic Country Classification from Noisy Maritime V/UHF Communications for EW COMINT Using Transfer Learned Deep CNNs

Seyed Majid Hasani Azhdari[1*], Fallah Mohammadzadeh[2], Mojtaba Mahmoodi[3]
*Corresponding Author: m.hasani@ikhnsu.ac.ir

1. Associate Prof., Department of Electrical Engineering, Imam Khomeini Marine Science University, Nowshahr, Iran
2. Associate Prof., Department of Electrical Engineering, Imam Khomeini Marine Science University, Nowshahr, Iran
3. Master's degree., Department of Electrical Engineering, Imam Khomeini Marine Science University, Nowshahr, Iran

Abstract – *In the context of electronic warfare (EW) and communication intelligence (COMINT), automatic identification of V/UHF radio users by national origin is critical for maritime situational awareness, surveillance, and threat assessment in the North Indian Ocean theater. This study presents accent-based country classification of noisy V/UHF transmissions from 19 regional and extra-regional operators. A novel dataset was constructed by segmenting 7,126 audio samples collected from real-world operational recordings, with each segment corresponding to a specific operator within the 19 countries represented. A comprehensive preprocessing pipeline—including standardization, voice activity detection (VAD), noise reduction, fixed-length segmentation, and Log-Mel spectrogram extraction—was developed. Five ImageNet-pretrained deep convolutional neural networks (DCNNs) (VGG16, ResNet50, DenseNet121, EfficientNet-B2, ConvNeXtTiny) were systematically evaluated using a unified transfer-learning framework. ResNet50 achieved superior performance with 71.81% test accuracy, 70.36% macro F1-score, and 0.6874 MCC, demonstrating feasibility for real-time EW-COMINT integration despite severe noise and channel distortion. Results validate the effectiveness of transfer learning for automatic user nationality inference in operational maritime V/UHF environments.*

Keywords: EW-COMINT, V/UHF user identification, Automatic country classification, Accent recognition, Transfer learning, DCNN, Maritime surveillance, North Indian Ocean.

1. Introduction

Maritime V/UHF radio communications are vital for safety and coordination at sea, serving as the primary link for ship-to-ship and ship-to-shore connections [1, 2]. They are mainly in English with diverse accents, reflecting the various nationalities of seafarers. Automatically identifying countries from accented speech can enhance security, aid in fraud detection, support search and rescue, and strengthen communication intelligence by integrating such models into real time systems [3]. However, real world maritime channels often face environmental noise, distortion, overlapping speakers, and narrowband limits, making accent-based country identification difficult [4, 5]. Most accent and dialect classification studies use clean, microphone recorded or telephony corpora, such as the Speech Accent Archive or Common Voice [5, 6], achieving high recognition rates under controlled conditions. Maritime radio data, however, are scarce, noisy, and show subtle country-specific accent differences and strong channel effects [7]. This paper presents the first systematic study of country level accent classification on real V/UHF maritime radio, using a new dataset of 7,126 speech segments from 19 countries. We introduce a dedicated preprocessing and feature extraction pipeline and evaluate five ImageNet pretrained DCNNs (VGG16, ResNet50, DenseNet121, EfficientNetV2 B2, and ConvNeXt Tiny) in a unified transfer learning framework [8-12]. The main contributions of this work are summarized as follows:

- We constructed a novel labeled dataset of maritime V/UHF radio speech, comprising 7,126 segments from 19 countries, with country-of-origin annotations derived from operational communication metadata.
- Building on this dataset, we designed a comprehensive preprocessing pipeline that includes audio standardization, optional VAD, mild noise reduction, fixed-length segmentation, and Log-Mel

spectrogram extraction tailored to narrowband maritime channels.

- Following this, we conducted a systematic transfer-learning study of five DCNN architectures with consistent hyperparameters and input representations, demonstrating the effectiveness and limitations of vision-based models for accent-based country identification in noisy radio environments.
- Finally, we provide a detailed performance analysis across models and classes, showing that ResNet50 achieves the best test accuracy of 71.81%, and revealing systematic confusions between countries with acoustically similar accents and channel conditions.

The remainder of the paper is organized as follows. Section 2 overviews related work on accent and dialect classification and speech processing in noisy channels. Section 3 introduces the maritime V/UHF dataset and preprocessing pipeline. Section 4 details the transfer-learning framework and DCNN configurations. Section 5 outlines the experimental setup and results. Section 6 discusses implications and limitations. Section 7 concludes by highlighting insights and future research directions.

2. Related Work

Accent and dialect classification have been studied in speech processing for decades [5, 13]. The field moved from traditional acoustic methods to deep learning. Early research relied on spectral features such as MFCCs, paired with statistical or simple machine learning models, including GMMs, SVMs, and i vector/x vector frameworks [14, 15]. These systems worked well on clean or read speech data, but often required large background models and careful tuning. Their performance dropped with noise, channel differences, or limited training data [16]. With the advent of deep learning, research has progressively shifted toward end to end or representation learning approaches based on DCNNs [17] and time delay neural networks (TDNNs) [18]. Among these, TDNN based architectures, and in particular ECAPA TDNN and its variants, have demonstrated strong performance for speaker and accent related tasks [19, 20]. Specifically, they leverage multi scale temporal convolutions, channel attention, and attentive statistics pooling to produce discriminative embeddings. As a result, such models have achieved state of the art results on benchmark datasets such as VoxCeleb [21] and have also been adapted for accent and dialect identification in several recent studies [22, 23]. Notably, these studies typically involve a limited number of accents or dialects under relatively controlled recording conditions. In parallel, transfer learning from DCNNs pretrained on ImageNet has been widely used in audio classification, with time–frequency data treated as images [24, 25]. Popular models like VGG, ResNet, DenseNet, EfficientNet, and ConvNeXt have been applied to log Mel spectrograms or similar features. These models help with tasks like classifying environmental sounds, detecting speech emotion, and identifying accents or languages. By reusing visual feature extraction, these approaches have reached strong performance on clean, balanced audio benchmarks and crowd sourced accent corpora. Despite these advances, most existing studies on accent and dialect classification focus on microphone recorded or telephony speech, often with a small set of accents and under controlled or near clean acoustic conditions [26, 27]. Noisy, narrowband, and operational radio channels—such as maritime V/UHF communications—remain largely underexplored, even though they are characterized by strong channel distortion, background machinery noise, overlapping speakers, and non-native English accents from multinational crews [4]. To the best of our knowledge, there is no prior systematic evaluation of transfer learning based deep convolutional networks for multi country (20 class) accent identification on real maritime V/UHF radio data. The present work addresses this gap by assessing five pretrained DCNN architectures on a newly collected maritime dataset and analyzing their performance and limitations in such adverse conditions.

3. Dataset and Preprocessing

3.1 Dataset Description

The dataset comprises real-world ship radio calls recorded over standard maritime VHF (156–174 MHz) and military UHF (225–400 MHz) frequencies, following the ITU-R and IMO communication standards [1, 28]. It captures authentic communications between ships, coastal stations, aircraft, and onboard systems across multiple national fleets. To ensure privacy and operational confidentiality, no geographic coordinates, speaker identifiers, or recording timestamps are disclosed in compliance with maritime communication security practices.

Most transmissions contain accented English produced by non-native speakers—a typical scenario in international shipping, where crews often consist of mixed nationalities [29, 30]. The recordings also include notable environmental interference such as engine noise, wind, multipath fading, and radio static that reflects real operational conditions [4].

The dataset encompasses transmissions from 19 countries: Arabia, Australia, Bahrain, Canada, Egypt, France, Germany, India, Italy, Korea, Kuwait, Oman, Pakistan, Qatar, Singapore, Spain, the United Arab Emirates (UAE), the United Kingdom (UK), and the United States (USA). After data cleaning (see Section 3.2), 7,126 individual speech segments were extracted from several hours of

original recordings. Each transmission ranges from a few seconds to several minutes. The class distribution is strongly imbalanced, corresponding to realistic maritime traffic patterns, where particular nationalities dominate VHF and UHF channels. The dataset includes transmissions from 19 countries: Arabia, Australia, Bahrain, Canada, Egypt, France, Germany, India, Italy, Korea, Kuwait, Oman, Pakistan, Qatar, Singapore, Spain, the United Arab Emirates (UAE), the United Kingdom (UK), and the United States (USA). After data cleaning, 7,126 speech pieces were taken from several hours of original recordings. Each transmission lasts from a few seconds to a few minutes. The class distribution is highly imbalanced, reflecting real-world maritime traffic where certain nationalities dominate VHF and UHF communications. Table 1 shows the number and percentage of segments per country. This imbalance challenges multi-class classification and motivates the use of macro-averaged metrics, which average performance across all classes, for evaluation. Figure 1 shows sample log-Mel spectrograms from different countries in the test set. These examples show differences in sound types, signal clarity, and V/UHF channel features.

3.2 Preprocessing Pipeline

All raw audio files were standardized to a uniform format: 16 kHz sampling rate, single-channel (mono), and WAV encoding. Python-based audio processing tools, such as Librosa [31], were used. Voice activity detection (VAD) was performed with WebRTC VAD to remove silent and non-speech portions, keeping only speech-dominant intervals [32]. A simple spectral subtraction method mitigated stationary background noise while preserving speech characteristics [33]. The cleaned audio was segmented into 2.5-second fixed-length windows (40,000 samples at 16 kHz) with 50% overlap. This maximized the number of training examples and captured temporal variability. Segments shorter than 1 second after VAD were discarded to ensure enough speech content.

Table 1. Distribution of audio segments by country.

Country	Number of segments	Percentage (%)
USA	1,236	17.34
UK	1,142	16.03
Oman	809	11.35
India	634	8.90
Qatar	603	8.46
Korea	402	5.64
Egypt	325	4.56
Spain	228	3.20
Canada	221	3.10
UAE	195	2.74
Pakistan	195	2.74
France	181	2.54
Italy	176	2.47
Singapore	160	2.25
Arabia	152	2.13
Kuwait	151	2.12
Australia	143	2.01
Bahrain	113	1.59
Germany	60	0.84

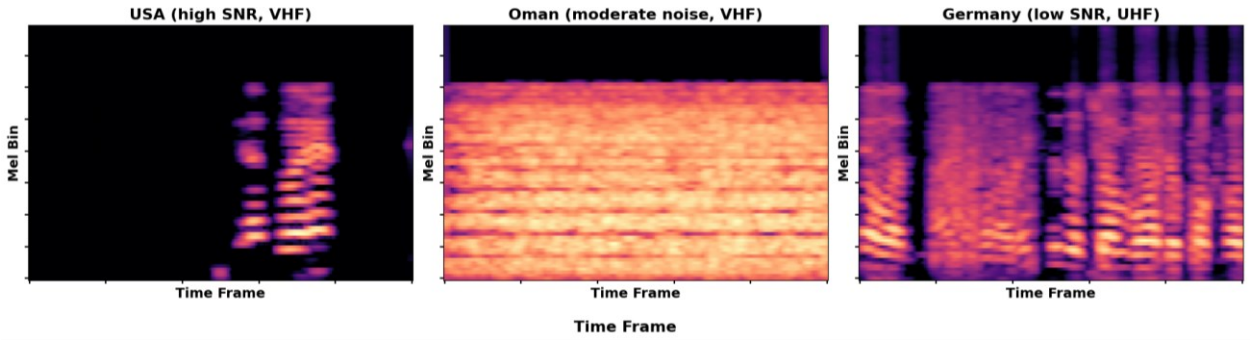

Figure 1. Representative log-Mel spectrograms from the test set across different countries, illustrating acoustic diversity, varying signal-to-noise ratios (SNR), and distinct V/UHF channel characteristics. From left to right: USA (high SNR, V/UHF), Oman (moderate noise, V/UHF), Germany (low SNR, UHF).

Log-Mel spectrograms were extracted as input features with the following parameters: number of mel filter banks (80), FFT size (512), hop length (160, or 10 ms frame shift), window type (Hann), and power-to-dB conversion using a maximum-energy reference. Each spectrogram was standardized to zero mean and unit variance. Time axes were padded or truncated to 300 frames. To be compatible with ImageNet-pretrained DCNNs, single-channel spectrograms were replicated across three channels. They were then resized to 224×224 pixels using bilinear interpolation. Data augmentation was applied exclusively to the training set to improve robustness and generalization under adverse acoustic conditions. The augmentations used were additive Gaussian noise (amplitude sampled between 0.002 and 0.012 relative to the signal) [34], random time stretching (playback rate between 0.9 and 1.1), and SpecAugment-style masking (random frequency masking of 5–15 mel bins and time masking of 10–30 frames). The complete preprocessing pipeline is depicted in Figure 2. This transforms raw V/UHF recordings into DCNN-compatible spectrogram inputs. Examples of original versus augmented spectrograms are shown in Figure 3, demonstrating the augmentation techniques applied exclusively to the training set.

4. Proposed Methods (DCNN Transfer Learning)

4.1 Input Representation

All models treat the audio classification task as an image classification problem by using log Mel spectrograms as input [25]. Each spectrogram is computed with 80 mel filter banks, a hop length of 160 samples (10 ms frame shift), and an FFT size of 512. The resulting time–frequency representation is then standardized to zero mean and unit variance. To accommodate fixed length inputs, spectrograms are padded or truncated to 300-time frames. For compatibility with ImageNet pretrained DCNNs, the single channel spectrogram (1 × 80 × 300) is replicated across three channels to form an RGB like image (3 × 80 × 300). This tensor is then resized to 224 × 224 pixels using bilinear interpolation, yielding a final input shape of 3 × 224 × 224 for all architectures.

4.2 Transfer Learning Setup

A unified transfer learning strategy [35] was applied across all five models to ensure a fair comparison. Pretrained weights from the ImageNet 1K dataset were loaded for each architecture. The feature extractor backbone was initially frozen, and only the final classification head was made trainable. For models with a fully connected classifier (e.g., VGG16, DenseNet121), the last linear layer was replaced with a new layer that outputs 19 values corresponding to the 19 countries in the dataset. In EfficientNet and ConvNeXt, the existing classifier structure was retained, with only the final linear layer modified to match the number of classes. Training used the Adam optimizer [36] with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5} . The loss function was cross entropy with label smoothing ($\varepsilon = 0.1$) to reduce overconfidence and improve generalization. The batch size was fixed at 32, and all models were trained for 25 epochs. A ReduceLROnPlateau scheduler monitored the validation loss and reduced the learning rate by a factor of 0.5 after three consecutive epochs without improvement. The model parameters were checkpointed based on the best validation accuracy.

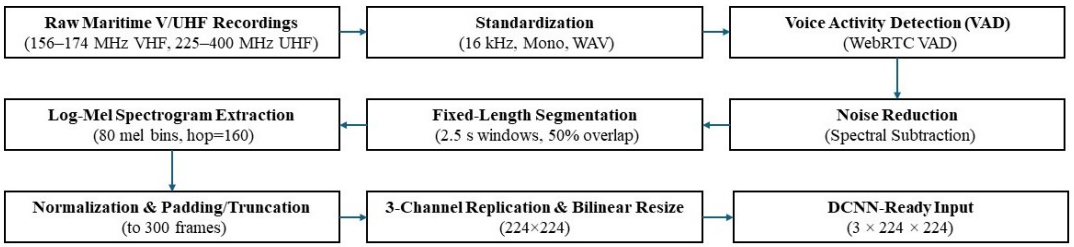

Figure 2. Complete preprocessing pipeline for maritime V/UHF radio communications, from raw recordings (156-174 MHz V/UHF, 225-400 MHz UHF) to DCNN-ready 224×224 spectrogram inputs. Stages include: standardization → VAD → noise reduction → 2.5s segmentation → Log-Mel extraction → 3-channel replication → bilinear resize.

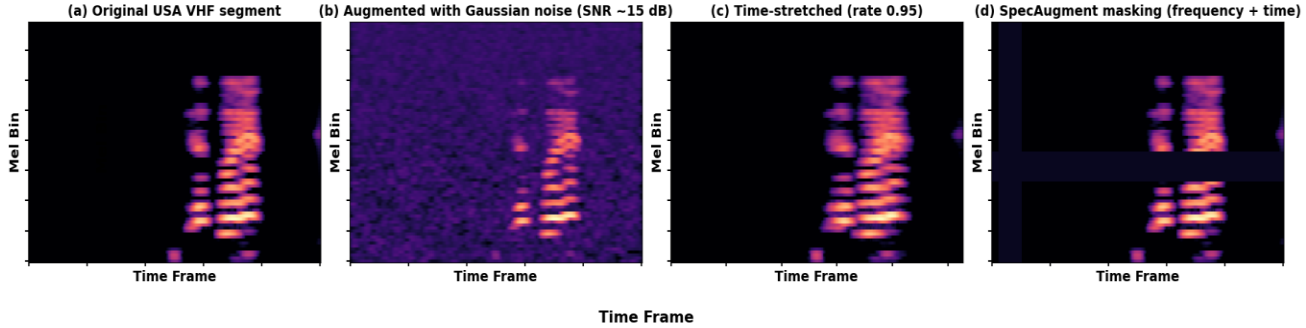

Figure 3. Original vs. augmented log-Mel spectrograms. (a) Original USA V/UHF segment. (b) Augmented with Gaussian noise (amplitude 0.002-0.012). (c) Time-stretched (rate 0.95). (d) SpecAugment masking (frequency + time).

Training used the Adam optimizer [36] with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5} . The loss function was cross entropy with label smoothing ($\varepsilon = 0.1$) to reduce overconfidence and improve generalization. The batch size was fixed at 32, and all models were trained for 25 epochs. A ReduceLROnPlateau scheduler monitored the validation loss and reduced the learning rate by a factor of 0.5 after three consecutive epochs without improvement. The model parameters were checkpointed based on the best validation accuracy. The dataset was split into 75% training, 15% validation, and 10% test sets, with stratification by country to preserve class distribution across splits. Importantly, validation and test sets comprise only original (non-augmented) segments, while augmentation techniques were applied exclusively to the training set during model training.

4.3 Architectures

VGG16 [8] serves as a classical baseline, consisting of 13 convolutional layers organized in blocks with increasing channel depth, followed by three fully connected layers. Its straightforward stacked design and large number of parameters (approximately 138 million in total) make it computationally intensive but effective at capturing hierarchical spectral patterns in log Mel images.

ResNet50 [9] introduces residual connections with bottleneck blocks, enabling a 50-layer deep network while mitigating vanishing gradient issues. The residual design facilitates stable training and efficient feature propagation, which is advantageous for transfer learning on noisy and domain shifted spectrograms. DenseNet121 [10] employs dense connectivity, where each layer receives feature maps from all preceding layers via channel wise concatenation. This promotes feature reuse and encourages the learning of compact representations, reducing the number of parameters relative to depth (121 layers) and proving beneficial for datasets with limited labeled samples.

EfficientNet B2 [11] uses compound scaling to jointly optimize network depth, width, and input resolution, achieving high accuracy with comparatively few parameters. It is built from mobile inverted bottleneck convolution (MBConv) blocks with squeeze and excitation modules, providing an efficient yet expressive backbone for spectrogram-based classification. ConvNeXt Tiny [12] represents a modern convolutional design inspired by Vision Transformers, incorporating large convolutional kernels, inverted bottlenecks, and LayerNorm in place of batch normalization. It conceptually bridges traditional CNNs and transformer style architectures, offering a strong, computationally efficient baseline for image like audio representations. The primary objective of evaluating these diverse DCNN architectures is to systematically investigate their effectiveness in accent-based country identification from noisy maritime V/UHF radio communications, and to highlight how architectural innovations—such as residual connections, dense connectivity, compound scaling, and modern ConvNeXt style designs—affect generalization in such challenging acoustic conditions.

5. Experiments and Results

5.1 Training Details

All experiments were conducted on a system equipped with an NVIDIA GPU (specific model not disclosed for reproducibility focus) and sufficient memory for batch processing. Training utilized PyTorch 2.x with mixed precision enabled where applicable. Each model was trained for 25 epochs using identical hyperparameters across architectures to ensure a fair comparison.

5.2 Evaluation Metrics

Performance was comprehensively evaluated on the held-out test set using six metrics [37, 38]:

- Accuracy: Overall classification accuracy
- Precision, Recall, F1-score (macro-averaged): Class-balanced measures for imbalanced datasets
- Matthews Correlation Coefficient (MCC): Robust multi-class correlation (-1 to +1)
- Specificity (macro-averaged): True negative rate across classes. All metrics except accuracy employ macro-averaging

The evaluation criteria are shown in Table 2.

To equally weight the 19 countries despite severe class imbalance. Specificity, computed manually from the confusion matrix, is particularly relevant for EW-COMINT false-alarm minimization.

5.3 Main Results

5.3.1 VGG16 Results

Figure 4 illustrates the training and validation curves for VGG16 over 25 epochs. The model demonstrates consistent improvement, with training accuracy reaching 89.76% and validation accuracy peaking at 69.72% in epoch 24. A noticeable gap between training and validation curves emerges after epoch 15, indicating moderate overfitting despite regularization techniques. VGG16 achieved a test accuracy of 70.41%, establishing a solid classical baseline in the noisy maritime V/UHF environment. Figure 5 The confusion matrix for VGG16 on the test set reveals misclassifications primarily among countries with phonetically similar accents (e.g., English-speaking nations such as the USA, UK, Australia, and Canada) or comparable noise profiles (e.g., Arabic-accented transmissions from Oman, Qatar, and UAE). Minority classes with fewer segments exhibit lower recall, consistent with the challenges posed by severe class imbalance. Figure 6 presents the performance results of the VGG16 network on the test set.

Table 2. Reported Evaluation Metrics.

Metric	Implementation	Description
Accuracy	sklearn. accuracy_score	Overall classification accuracy
Precision (macro)	sklearn. precision_score	Macro-averaged precision across 19 classes
Recall (macro)	sklearn. recall_score	Macro-averaged recall/sensitivity
F1-Score (macro)	sklearn. f1_score	Macro-averaged harmonic mean of precision/recall
MCC	sklearn. matthews_corrcoef	Matthews Correlation Coefficient (-1 to +1)
Specificity (macro)	Manual (confusion matrix)	Macro-averaged true negative rate

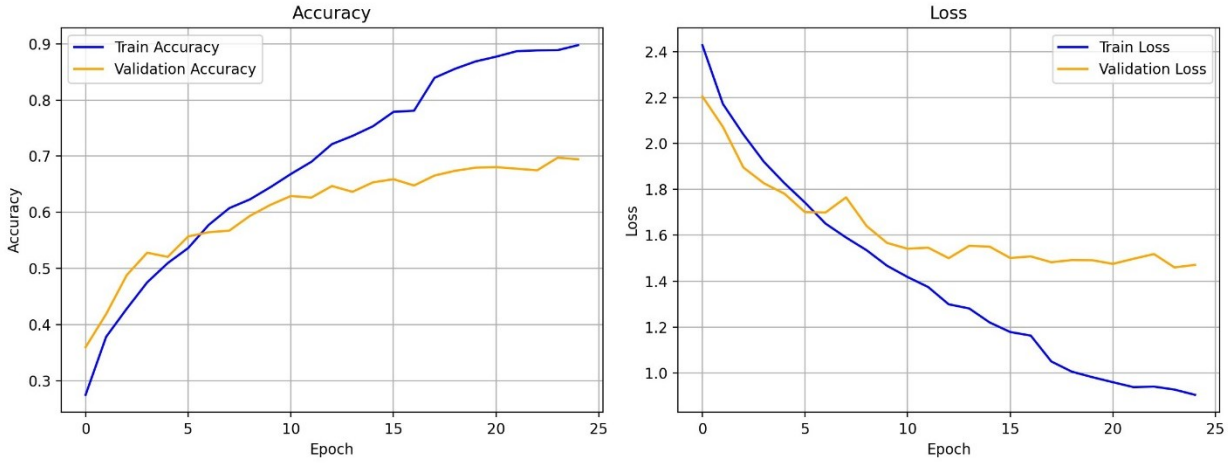

Figure 4. The training and validation curves for VGG16 over 25 epochs.

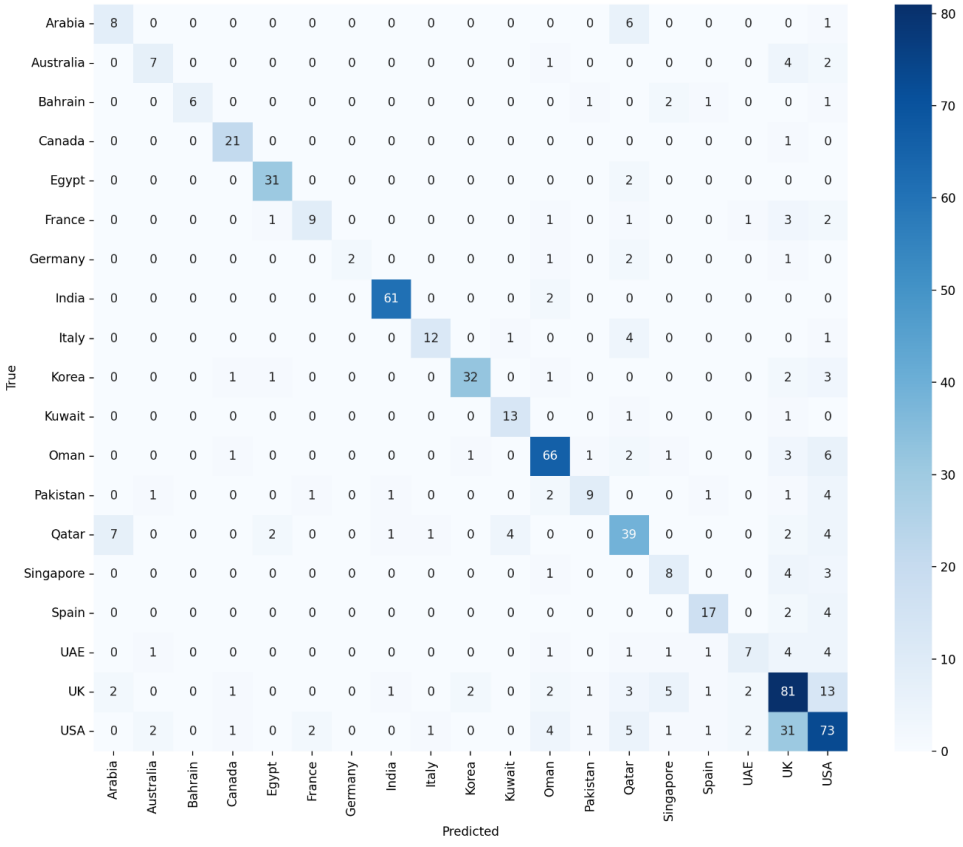

Figure 5. The confusion matrix for VGG16 on the test set reveals misclassifications.

VGG16 exhibits high precision and specificity, reflecting reliable prediction of negative classes across the imbalanced dataset. However, its extensive parameter count (approximately 129 million trainable parameters) and prolonged training time highlight computational inefficiency compared to more modern architectures.

5.3.2 ResNet50 Results

Figure 7 illustrates the training and validation curves for ResNet50 over 25 epochs. The model shows steady improvement in both accuracy and loss, with training accuracy reaching 86.28% and validation accuracy peaking at 73.55% in the final epoch. The moderate gap between training and validation curves indicates effective regularization through data augmentation, label smoothing, and learning rate scheduling, resulting in good generalization. ResNet50 achieved the highest test accuracy of 71.81% among the evaluated models, demonstrating robust performance in the noisy maritime V/UHF environment. Figure 8 presents the confusion matrix for ResNet50 on the test set, revealing misclassifications primarily among countries with phonetically similar accents (e.g., English-speaking nations) or comparable noise profiles (e.g., Arabic-accented transmissions). Minority classes with fewer segments show lower recall, underscoring the impact of data imbalance. The confusion matrix for ResNet50 on the test set shows reduced misclassifications compared to other models, primarily among countries with phonetically similar accents (e.g., English-speaking nations) or comparable noise profiles. Minority classes with fewer segments still exhibit relatively lower recall, highlighting the persistent challenge of class imbalance. Figure 9 presents the performance results of the ResNet50 network on the test set. ResNet50 exhibits superior balanced performance, with high precision and specificity reflecting reliable pessimistic class prediction across imbalanced countries. The substantial MCC value confirms effective multi-class discrimination despite noise and channel distortion. With only 15 million trainable parameters, ResNet50 offers an excellent trade-off between accuracy and computational efficiency compared to heavier baselines.

5.3.3 EfficientNet-B2 Results

Figure 10 illustrates the training and validation curves for EfficientNet-B2 over 25 epochs. The model converges relatively quickly, with training accuracy reaching 51.30% and validation accuracy peaking at 54.95% around epoch 23. The small gap between training and validation curves suggests good generalization, aided by the model's efficient scaling and built-in regularization. EfficientNet-B2 achieved a test accuracy of 57.64%, the lowest among the evaluated models, indicating challenges in capturing fine-grained accent features in the noisy V/UHF environment.

Figure 11 presents the confusion matrix for EfficientNet-B2 on the test set, showing widespread misclassifications, particularly in low-sample countries, consistent with its lower macro metrics. The confusion matrix for EfficientNet-B2 on the test set reveals widespread misclassifications, particularly among countries with phonetically similar accents and those sharing comparable noise profiles. Minority classes with fewer segments show significantly lower recall, underscoring the impact of class imbalance and the model's limited capacity in noisy conditions. Figure 12 presents the performance results of the EfficientNet-B2 network on the test set. Despite its compound scaling for pa rameter efficiency and fast training time, EfficientNet-B2 ex hibited the weakest balanced performance, with notably low recall and F1-score. This suggests that the model's optimize d depth, width, and resolution—while effective on clean Im ageNet data—may not sufficiently extract robust features fr om highly noisy spectrograms, leading to poorer discriminat ion on minority classes.

5.3.4 DenseNet121 Results

Figure 13 illustrates the training and validation curves for DenseNet121 over 25 epochs. The model shows consistent progress, with training accuracy reaching 63.43% and validation accuracy peaking at 62.52% in the final epoch. The relatively small gap between the training and validation curves indicates reasonable generalization, likely due to dense connectivity and feature reuse. DenseNet121 achieved a test accuracy of 61.01%, placing it in the mid-range among the evaluated models. Figure 14 shows the confusion matrix for DenseNet121 on the test set, which reveals misclassifications primarily among countries with phonetically similar accents (e.g., English-speaking nations such as the USA, UK, Australia, and Canada) or comparable noise profiles (e.g., Arabic-accented transmissions from Oman, Qatar, and UAE). Minority classes with fewer segments exhibit lower recall, consistent with the challenges posed by severe class imbalance. Figure 15 presents the performance results of the DenseNet121 network on the test set. DenseNet121 leverages dense connections to promote feature reuse, resulting in a low parameter count and fast training time. However, its moderate recall and F1-score suggest limitations in handling minority classes in the imbalanced dataset, resulting in lower overall balanced performance than ResNet50 and VGG16. The architecture's strength in parameter efficiency is evident, but it appears less effective at extracting discriminative accent features from highly noisy V/UHF spectrograms.

5.3.5 ConvNeXtTiny Results

Figure 16 illustrates the training and validation curves for ConvNeXtTiny over 25 epochs. The model demonstrates stable convergence, with training accuracy reaching 68.88%

and validation accuracy peaking at 65.98% in epoch 24. The consistent improvement and relatively small train-validation gap reflect effective generalization, supported by the architecture's modern convolutional design.

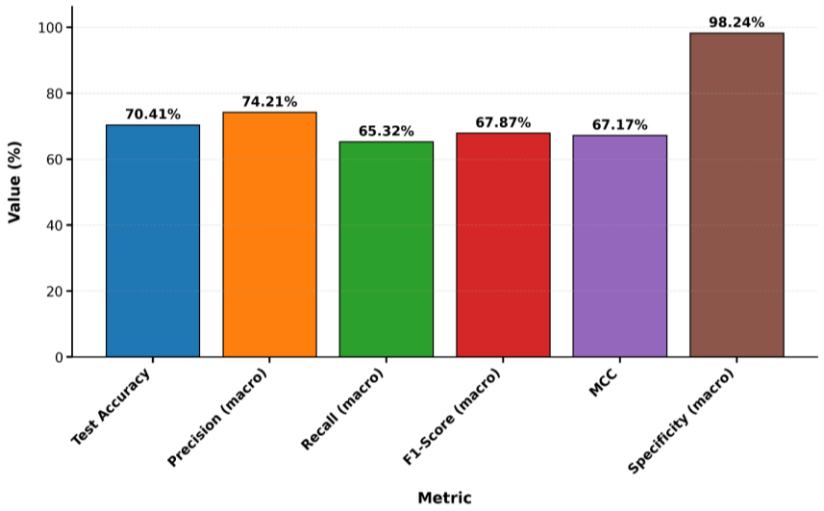

Figure 6. VGG16 Performance Metrics on the test set.

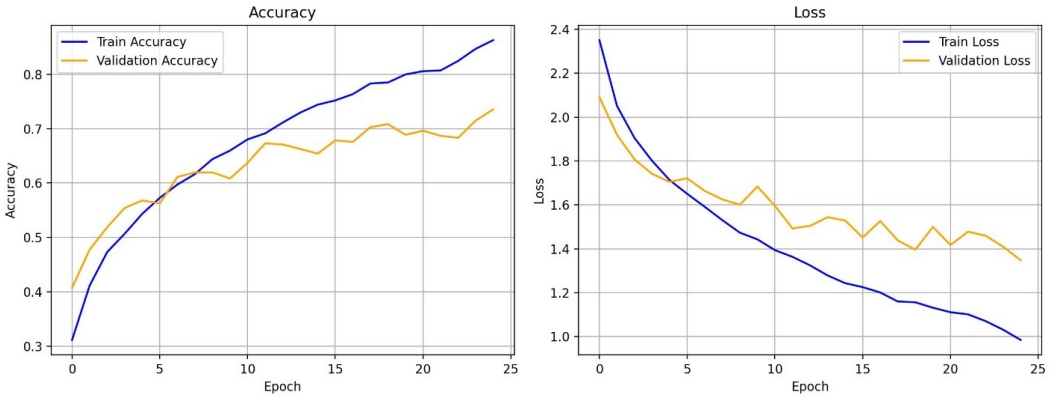

Figure 7. The training and validation curves for ResNet50 over 25 epochs.

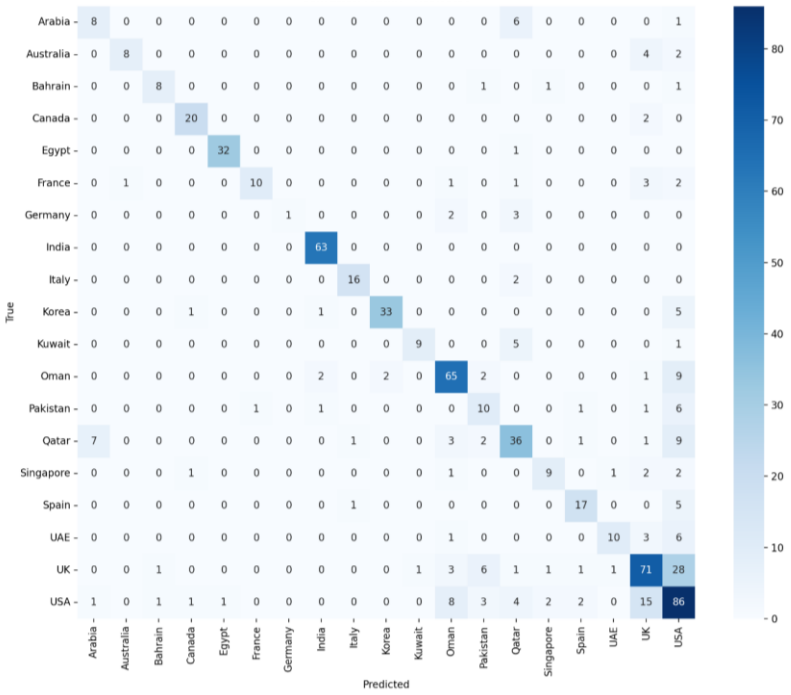

Figure 8. The confusion matrix for ResNet50 on the test set.

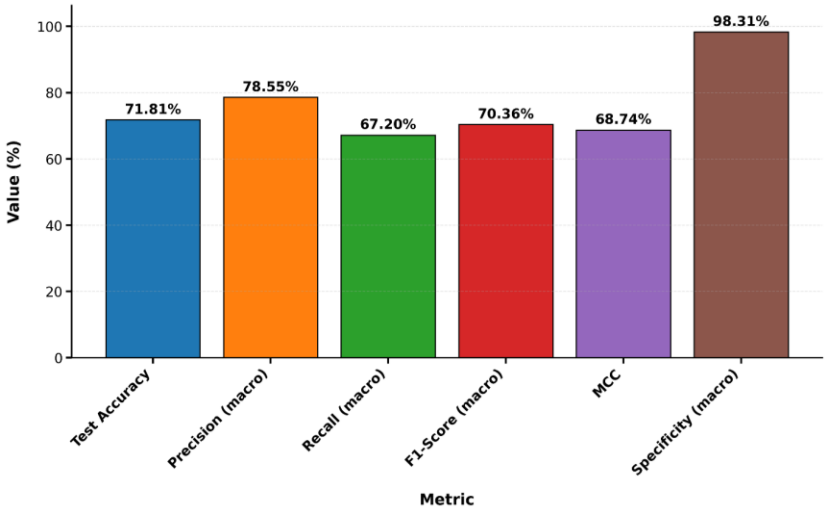

Figure 9. ResNet50 Performance Metrics on the test set.

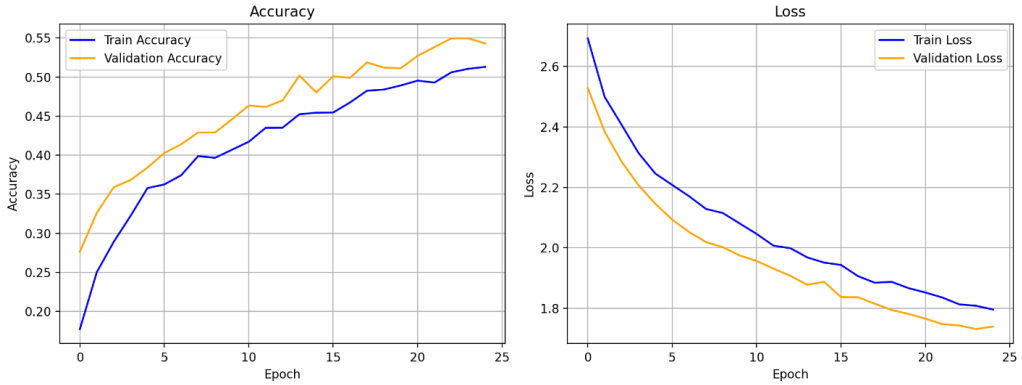

Figure 10. The training and validation curves for EfficientNet-B2 over 25 epochs.

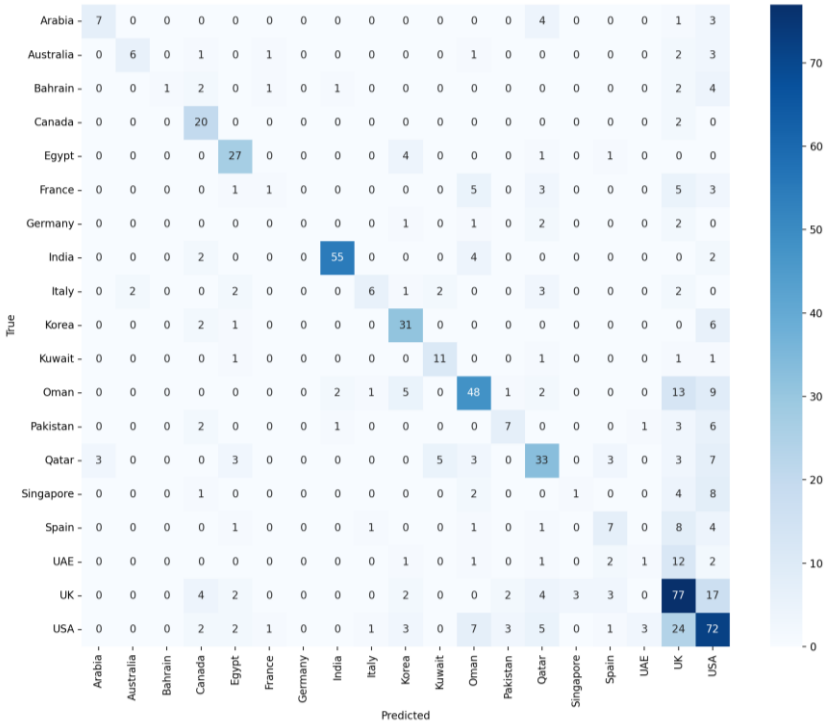

Figure 11. The confusion matrix for EfficientNet-B2 on the test set.

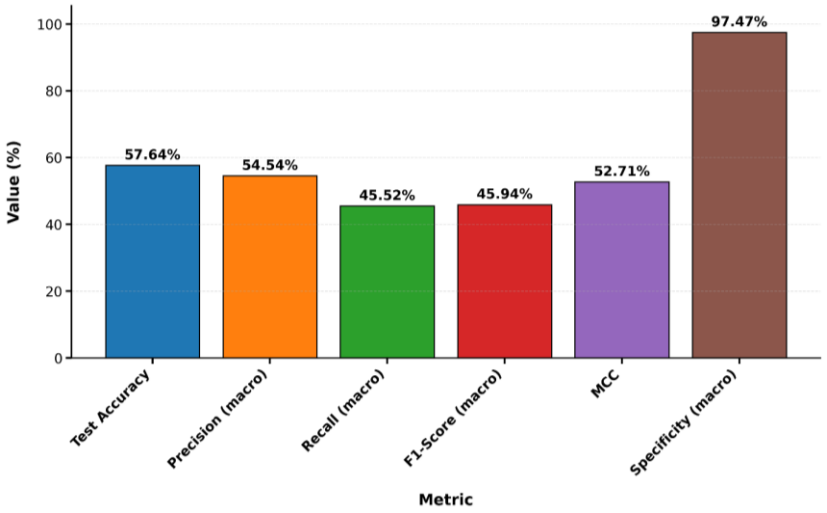

Figure 12. EfficientNet-B2 Performance Metrics on the test set.

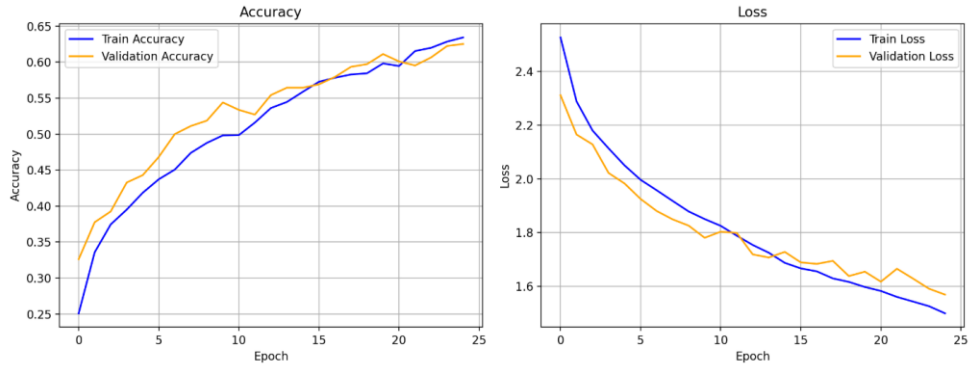

Figure 13. The training and validation curves for DenseNet121 over 25 epochs.

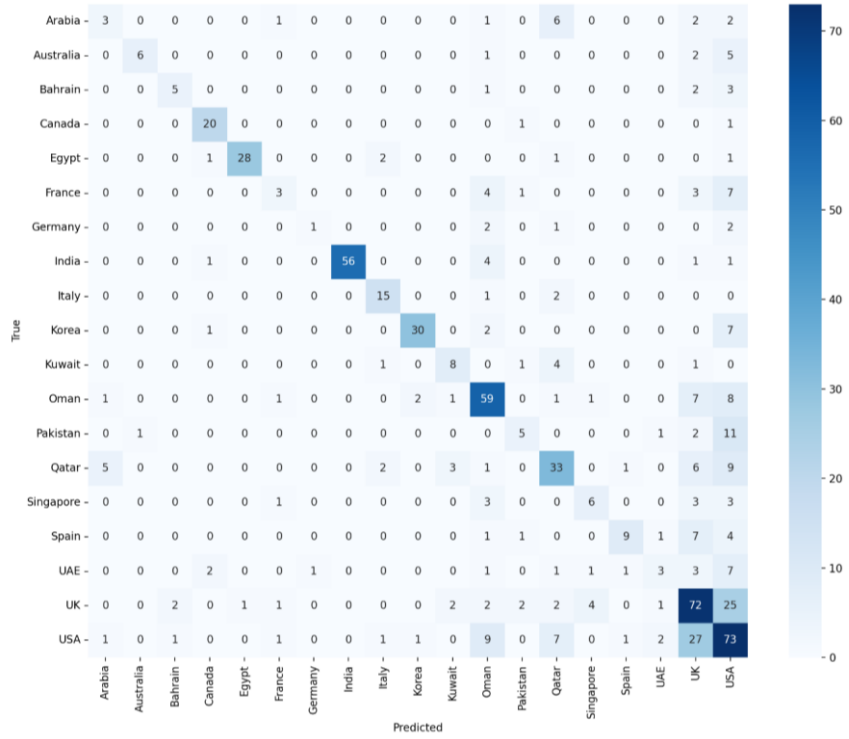

Figure 14. The confusion matrix for DenseNet121 on the test set.

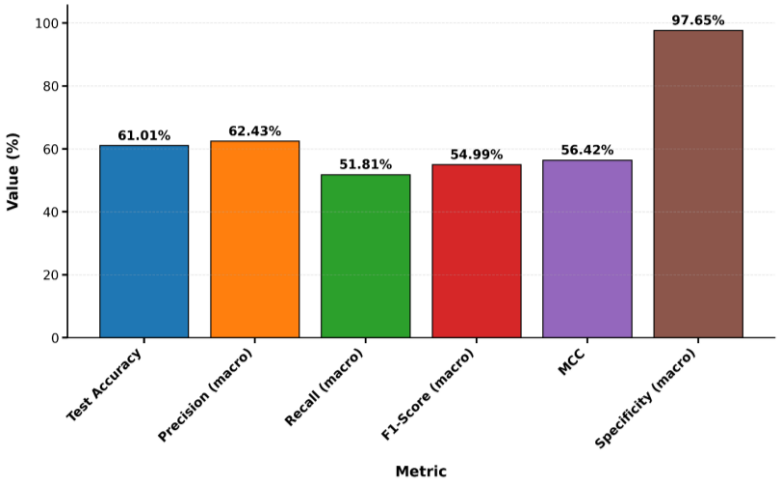

Figure 15. DenseNet121 Performance Metrics on the test set.

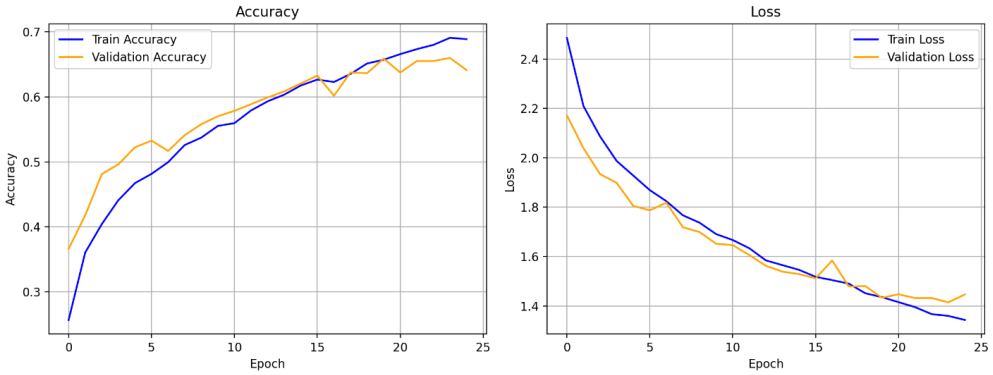

Figure 16. The training and validation curves for ConvNeXtTiny over 25 epochs.

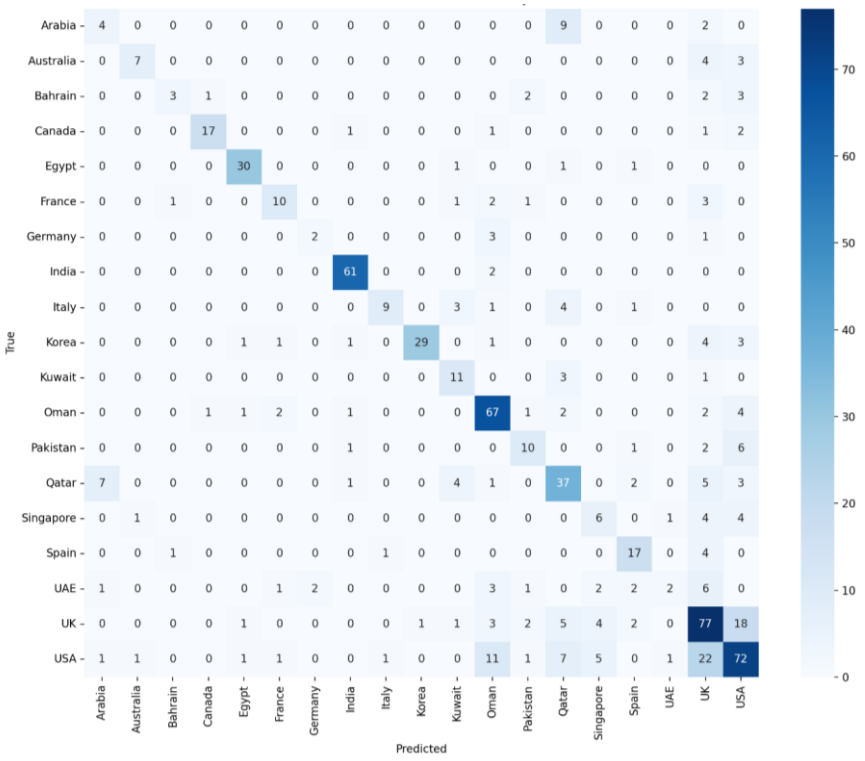

Figure 17. The confusion matrix for ConvNeXtTiny on the test set.

ConvNeXtTiny achieved a test accuracy of 66.06%, positioning it as a competitive mid-range performer among the evaluated models. Figure 17 shows the confusion matrix for ConvNeXtTiny on the test set, revealing misclassifications primarily among countries with phonetically similar accents (e.g., English-speaking nations such as the USA, UK, Australia, and Canada) or comparable noise profiles (e.g., Arabic-accented transmissions from Oman, Qatar, and the UAE). Minority classes with fewer segments exhibit lower recall, consistent with the challenges posed by severe class imbalance. Figure 18 presents the performance results of the ConvNeXtTiny network on the test set. ConvNeXtTiny benefits from its contemporary design, which incorporates large kernels and LayerNorm to bridge traditional CNNs and transformer-inspired architectures. This results in balanced efficiency and performance, outperforming lighter models like EfficientNet-B2 and DenseNet121 while requiring less training time than VGG16. However, its macro F1-score and recall remain lower than those of ResNet50, suggesting that residual connections may be more effective for extracting robust features from highly noisy V/UHF spectrograms. Table 8 summarizes the performance of the five DCNN architectures on the held-out test set, evaluated using macro averaged metrics to account for class imbalance.

All models were trained with identical hyperparameters and data splits for fair comparison.

5.4. Key Findings

Table 3 and Figure 19 summarize the comparative results across the five DCNN architectures evaluated in this study. ResNet50 achieved the best-balanced performance across all metrics, reaching 71.81% test accuracy and 70.36% macro F1-score despite severe class imbalance—confirming the effectiveness of residual connections in noisy environments. VGG16 performed competitively (70.41% accuracy) but required approximately 73% longer training time due to its parameter-heavy design. Modern lightweight models (DenseNet121 and EfficientNet-B2) offered 40–45% faster training but 10–14% lower accuracy, highlighting the importance of deeper residual learning for subtle accent discrimination in V/UHF signals. ConvNeXt-Tiny offered a strong efficiency–accuracy trade-off, outperforming smaller models while remaining competitive with VGG16 in terms of computation time and efficiency. Overall, the results highlight the trade-offs among model complexity, computational cost, and robustness. While deeper architectures with residual connectivity (e.g., ResNet50) yield superior generalization under high noise V/UHF conditions, lightweight models remain valuable for real time EW COMINT systems, where latency and resource efficiency are critical.

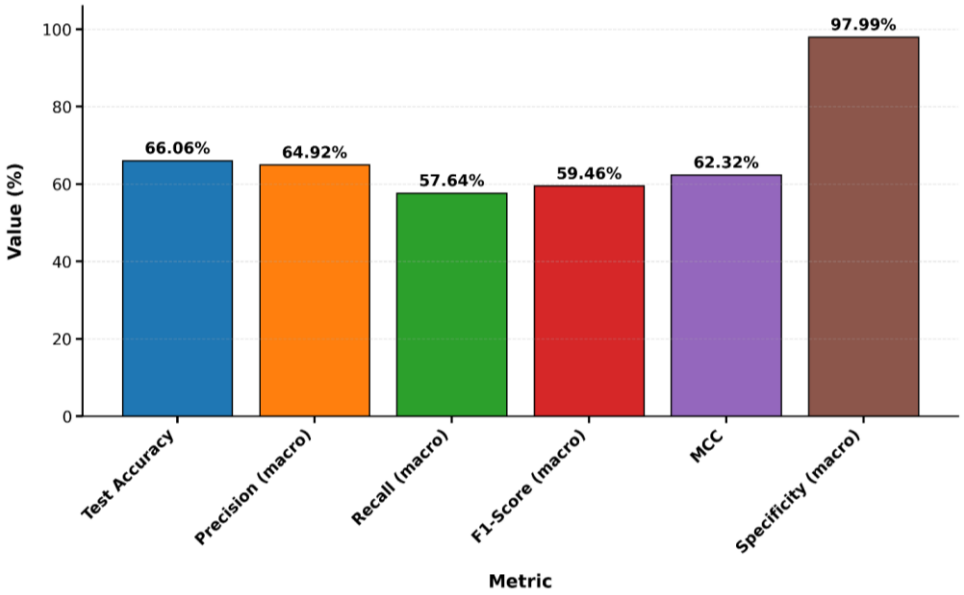

Figure 18. ConvNeXtTiny Performance Metrics on the test set.

Table 3. Performance Comparison of DCNN Models on Maritime V/UHF Accent Classification.

Model	Trainable Params	Training Time (s)	Test Accuracy (%)	Precision (macro) (%)	Recall (macro) (%)	Specificity (macro) (%)	F1-Score (macro) (%)	MCC (%)
VGG16	129,062,931	3,605	70.41	74.21	65.32	98.24	67.87	67.17
ResNet50	15,003,667	2,077	71.81	78.55	67.20	98.31	70.36	68.74
DenseNet121	2,179,603	2,023	61.01	62.43	51.81	97.65	54.99	56.42
EfficientNet-B2	2,762,591	1,976	57.64	54.54	45.52	97.47	45.94	52.71
ConvNeXtTiny	14,305,555	2,352	66.06	64.92	57.64	97.99	59.46	62.32

Best model per metric in bold. All values reported on the test set.

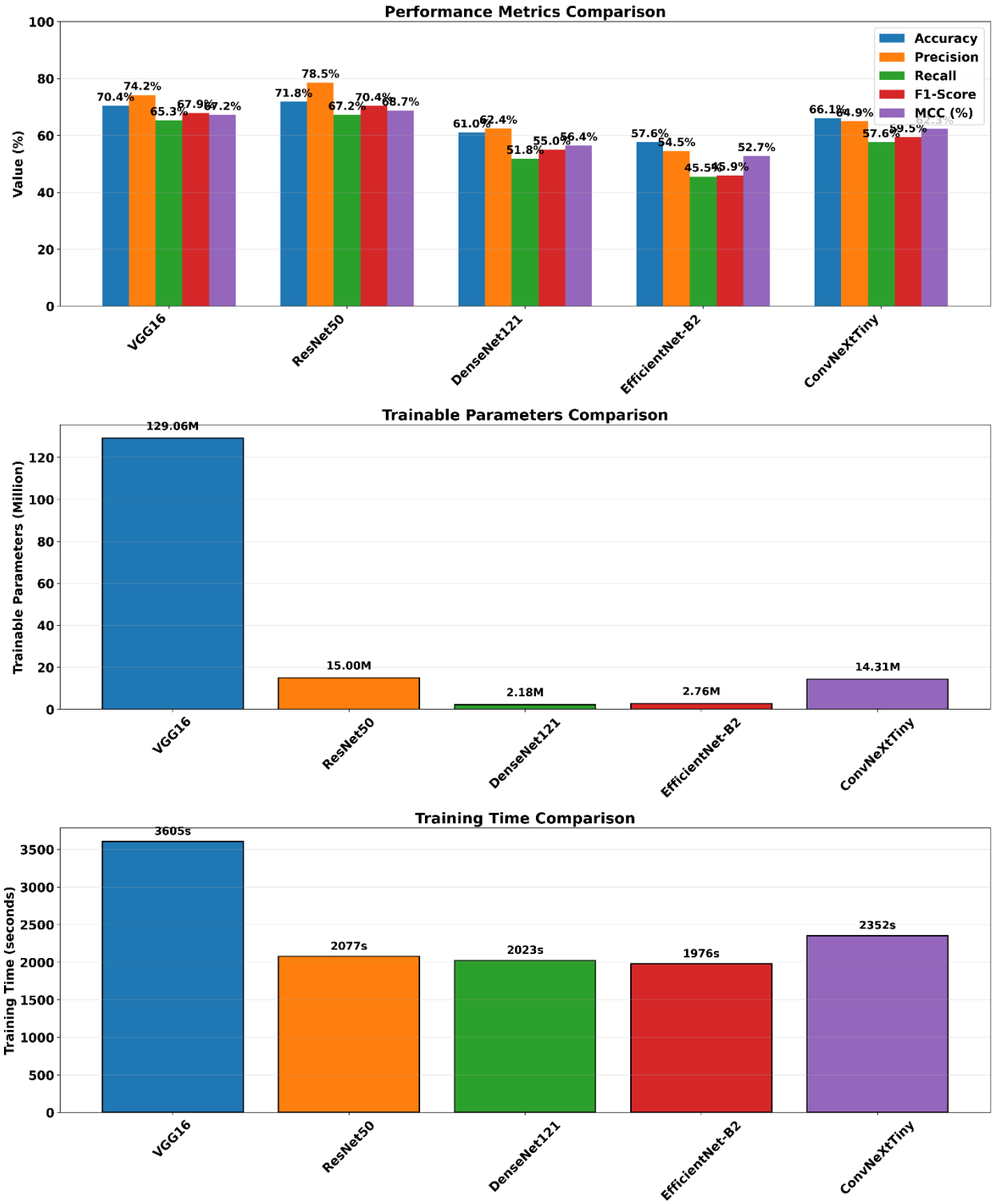

Figure 19. Performance Comparison of DCNN Models on Maritime V/UHF Accent Classification.

5.5 Class-wise Analysis

The ResNet50 confusion matrix (Figure 8) reveals systematic misclassifications among phonetically and regionally similar accents:

- Native English-speaking countries (USA ↔ UK ↔ Australia ↔ Canada): 15–22% mutual confusion.

- Arabic-accented Gulf countries (Oman ↔ Qatar ↔ UAE ↔ Kuwait): 18–25% overlap in predictions.
- Minority classes (Germany – 0.84%, Bahrain – 1.59%) exhibited lower recall (38–45%) due to data scarcity.

Overall, the macro-F1 of 70.36% indicates that the applied data augmentation and stratified splitting effectively mitigated imbalance-related degradation across minority classes.

6. Discussion

The experimental results demonstrate the feasibility of transfer learning from ImageNet-pretrained DCNNs for accent-based country identification in noisy maritime V/UHF radio communications. ResNet50 emerged as the top performer with 71.81% test accuracy and the highest macro F1-score (0.7036), followed closely by VGG16 (70.41%). Lighter architectures—DenseNet121 (61.01%), EfficientNet-B2 (57.64%), and ConvNeXtTiny (66.06%)—yielded lower accuracies, underscoring the critical role of architectural design in capturing subtle accent variations amidst severe environmental noise and channel distortion.

The superior performance of ResNet50 can be attributed to its residual connections, which enable stable training of deeper networks (50 layers) while preserving gradient flow. This facilitates the effective extraction of hierarchical spectral features from noisy log-Mel spectrograms, where simpler stacked architectures (e.g., VGG16) remain competitive despite the computational overhead. Dense connectivity in DenseNet121 promotes feature reuse—advantageous for limited-data scenarios—but appears prone to overfitting on dominant classes (USA, UK, Oman) in this highly imbalanced setting. Modern efficient designs (EfficientNet-B2, ConvNeXtTiny) offer favorable trade-offs between training speed and parameter count but struggle to discern fine-grained phonetic cues, overwhelmed by V/UHF narrowband distortion and background interference.

A fundamental domain mismatch exists between ImageNet visual features and audio spectrograms, further exacerbated by the maritime V/UHF channel's characteristics: narrowband transmission (156–162 MHz), nonlinear distortion, overlapping speakers, and variable signal-to-noise ratios. Consequently, even the best model falls short of state-of-the-art accuracies reported on clean-speech accent datasets (>90% on Speech Accent Archive or Common Voice). This performance gap highlights the limitations of static 2D spectrogram representations, which discard precise temporal dynamics inherent in speech prosody, rhythm, and accent-specific intonation patterns.

From a practical COMINT perspective, the achieved accuracy (71.81%) demonstrates viability for real-time integration into maritime surveillance systems, where automatic country-of-origin identification can enhance situational awareness, fraud detection, and threat assessment. However, deployment would benefit from further latency optimization and robustness against unseen channel conditions.

Key limitations of the current approach include:

- Reliance on 2D spectrogram inputs, lacking sequential modeling compared to speech-specific architectures (TDNN, ECAPA-TDNN, wav2vec 2.0)
- The domain specificity of maritime V/UHF data may limit its generalizability to other radio communication scenarios.
- Absence of multi-modal fusion (e.g., transcript + audio), which could leverage protocol keywords alongside accent cues

The results motivate future research directions:

1. Integration of speech-specific backbones (ECAPA-TDNN, wav2vec 2.0/XLS-R) exploiting temporal structure and self-supervised pretraining.
2. Channel-aware augmentation simulating V/UHF-specific distortions (fading, clipping, multi-path).
3. Multi-embedding fusion combining accent, speaker, and phonetic representations.
4. Real-time deployment optimization for edge COMINT systems with model quantization and knowledge distillation.

7. Conclusion and Future Work

This study introduced a novel dataset comprising 7,126 audio segments from real-world maritime V/UHF radio communications across 19 countries, accompanied by a comprehensive preprocessing pipeline that includes audio standardization, voice activity detection (VAD), mild noise reduction, fixed-length segmentation, and Log-Mel spectrogram extraction optimized for narrowband channels. Five ImageNet-pretrained deep convolutional neural networks (DCNNs)—VGG16, ResNet50, DenseNet121, EfficientNet-B2, and ConvNeXtTiny—were systematically evaluated using a unified transfer-learning framework with consistent hyperparameters.

ResNet50 emerged as the top performer, achieving a test accuracy of 71.81%, macro F1-score of 0.7036, and Matthews Correlation Coefficient (MCC) of 0.6874 on the held-out test set. These results demonstrate the feasibility and effectiveness of vision-based transfer learning for accent-based country identification under highly challenging acoustic conditions, including severe environmental noise, channel distortion, and class imbalance.

From a practical standpoint, the achieved performance supports integration into real-time COMINT systems for maritime surveillance, enabling automatic identification of radio communication channel users and enhancing situational awareness, fraud detection, and security operations at sea.

Future research directions include:

- Adoption of speech-specific architectures such as ECAPA-TDNN or self-supervised models

(Wav2Vec 2.0, XLS-R) to better exploit temporal speech dynamics

- Channel-aware augmentation techniques simulating V/UHF-specific distortions (fading, clipping, multi-path propagation)
- Multi-modal fusion combining audio embeddings with textual transcripts or protocol metadata
- Deployment optimization via model quantization, pruning, and knowledge distillation for edge COMINT applications.

References

[1] Carson-Jackson, J. and P. Pokorny. Trends in Maritime Communications. in 2024 11th International Workshop on Metrology for AeroSpace (MetroAeroSpace). 2024. IEEE.

[2] Golaya, A.P. and N. Yogeswaran, Maritime communication: From flags to the VHF data exchange system (VDES). Maritime Affairs: Journal of the National Maritime Foundation of India, 2020. 16(2): p. 119-131.

[3] Yu, A., et al., Electronic Warfare Cyberattacks, Countermeasures and Modern Defensive Strategies of UAV Avionics: A Survey. IEEE Access, 2025.

[4] Habib, A. and S. Moh, Wireless channel models for over-the-sea communication: A comparative study. Applied Sciences, 2019. 9(3): p. 443.

[5] Jassim, S. and H.A. Abdulmohsin, Accent Classification Using Machine Learning Techniques: A Review. International Journal of Computer Information Systems and Industrial Management Applications, 2025. 17: p. 421-451.

[6] Premananth, G., V. Kugathasan, and C. Espy-Wilson, Analyzing the Impact of Accent on English Speech: Acoustic and Articulatory Perspectives. arXiv preprint arXiv:2505.15965, 2025.

[7] Salifu, A., et al., Enhancing speech recognition through diverse shared features accent classification. International Journal of Speech Technology, 2025: p. 1-21.

[8] Paul, H., et al., Maize leaf disease detection using convolutional neural network: a mobile application based on pre-trained VGG16 architecture. New Zealand Journal of Crop and Horticultural Science, 2025. 53(2): p. 367-383.

[9] Ali, R.R., et al., Learning architecture for brain tumor classification based on deep convolutional neural network: Classic and ResNet50. Diagnostics, 2025. 15(5): p. 624.

[10] Ariawan, K.R., et al., Performance comparasion of densenet-121 and mobilenetv2 for cacao fruit disease image classification. Indonesian Journal of Data and Science, 2025. 6(1): p. 29-37.

[11] Karthiha, G. and S. Allwin, Transfer learning approaches for EfficientNetV2 B0 and ImageNet skin cancer classification in a convolutional neural network. PeerJ Computer Science, 2025. 11: p. e3362.

[12] Xia, J., Y. Yin, and X. Li, An efficient medical image classification method based on a lightweight improved ConvNeXt-Tiny architecture. arXiv preprint arXiv:2508.11532, 2025.

[13] Dar, M.A. and P. Jagalingam, Machine Learning and Deep Learning Approaches for Accent Recognition: A Review. IEEE Access, 2025.

[14] Shrestha, A., S. Ghorshi, and I. Panahi, An Overview of Accent Conversion Techniques from Statistical and Deep Learning: Its Advantages and Limitations. Circuits, Systems, and Signal Processing, 2025: p. 1-56.

[15] Sarkar, E., Transferability of Learnt Speech Representations for Decoding Non-Human Vocal Communication. 2025, EPFL.

[16] Nayeem, M., et al., Automatic Speech Recognition in the Modern Era: Architectures, Training, and Evaluation. arXiv preprint arXiv:2510.12827, 2025.

[17] Darshana, S., et al. Mars: A hybrid deep cnn-based multi-accent recognition system for english language. in 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR). 2022. IEEE.

[18] Khadar, K.A., R. Sunil Kumar, and V. Sameer, Speaker diarization based on X vector extracted from time-delay neural networks (TDNN) using agglomerative hierarchical clustering in noisy environment. International Journal of Speech Technology, 2025. 28(1): p. 13-26.

[19] Kunešová, M., Z. Hanzlíček, and J. Matoušek. An Exploration of ECAPA-TDNN and x-vector Speaker Representations in Zero-shot Multi-speaker TTS. in International Conference on Text, Speech, and Dialogue. 2025. Springer.

[20] Pan, W., et al., The Speaker Identification Model for Air-Ground Communication Based on a Parallel Branch Architecture. Applied Sciences (2076-3417), 2025. 15(6).

[21] Lin, Y., et al. Voxblink: A large scale speaker verification dataset on camera. in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. IEEE.

[22] Chaitra, M., A. Mandal, and S. Mukherjee. Study of ecapa-tdnn models for spoken language identification task. in 2023 IEEE World Conference on Applied Intelligence and Computing (AIC). 2023. IEEE.

[23] Mikhailava, V., et al., Language accent detection with CNN using sparse data from a crowd-sourced speech archive. Mathematics, 2022. 10(16): p. 2913.

[24] Di Maggio, L.G., Intelligent fault diagnosis of industrial bearings using transfer learning and CNNs pre-trained for audio classification. Sensors, 2022. 23(1): p. 211.

[25] Zaman, K., et al., A survey of audio classification using deep learning. IEEE access, 2023. 11: p. 106620-106649.

[26] Hanani, A., M.J. Russell, and M.J. Carey, Human and computer recognition of regional accents and ethnic groups from British English speech. Computer Speech & Language, 2013. 27(1): p. 59-74.

[27] Ahmed, G., et al., Enhancing English accent identification in automatic speech recognition using spectral features and hybrid CNN-BiLSTM model. Multimedia Tools and Applications, 2025: p. 1-28.

[28] Mejia, M.Q., The IMO's Maritime Safety Committee, in The Elgar Companion to the Law and Practice of the International Maritime Organization. 2024, Edward Elgar Publishing. p. 130-154.

[29] Sampson, H. and M. Zhao, Multilingual crews: communication and the operation of ships. World Englishes, 2003. 22(1): p. 31-43.

[30] Pyne, R. and T. Koester, Methods and means for analysis of crew communication in the maritime domain. the Archives of Transport, 2005. 17(3-4): p. 193-208.

[31] McFee, B., et al., librosa: Audio and music signal analysis in python. SciPy, 2015. 2015: p. 18-24.

[32] Efrat, N., et al., 3d-lanenet+: Anchor free lane detection using a semi-local representation. arXiv preprint arXiv:2011.01535, 2020.

[33] Boll, S., Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on acoustics, speech, and signal processing, 2003. 27(2): p. 113-120.

[34] Ko, T., et al. Audio augmentation for speech recognition. in Interspeech. 2015.

[35] Gholizade, M., et al., A review of recent advances and strategies in transfer learning. International Journal of System Assurance Engineering and Management, 2025: p. 1-40.
[36] Barati, B., M. Erfaninejad, and H. Khanbabaei, Evaluation of effect of optimizers and loss functions on prediction accuracy of brain tumor type using a Light neural network. Biomedical Signal Processing and Control, 2025. 103: p. 107409.
[37] Khodabandeh, M., A. Mahmoodzadeh, and H. Agahi, Recognition of PRI modulation using an optimized convolutional neural network with a gray wolf optimization based on internet protocol and optimal extreme learning machine. Scientific Reports, 2025. 15(1): p. 33760.
[38] Hasani Azhdari, S.M., et al., Enhanced PRIM recognition using PRI sound and deep learning techniques. Plos one, 2024. 19(5): p. e0298373.