Development of a Robust Neural Network-Based VAD System under Low Signal-to-Noise Ratio Conditions

Aigul Kulakayeva
Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University, Almaty, Kazakhstan
a.kulakayeva@iitu.edu.kz

Bekbolat Medetov
Department of Radio Engineering, Electronics and Telecommunications, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
bm02@mail.ru

Ainur Zhetpisbayeva
Department of Radio Engineering, Electronics and Telecommunications, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
aigulji@mail.ru

Aigul Nurlankyzy
Department of Electronics, Telecommunications and Space Technologies, Satbayev University, Almaty, Kazakhstan | Department of Cybersecurity, International Information Technology University, Almaty, Kazakhstan
nurlankyzyaigulya@gmail.com (corresponding author)

Received: 20 September 2025 | Revised: 10 October 2025 | Accepted: 21 October 2025

 DOI: https://doi.org/10.48084/etasr.14960

ABSTRACT

This study investigates the problem of developing and evaluating robust Voice Activity Detection (VAD) systems under low Signal-to-Noise Ratio (SNR) conditions, which presents a significant challenge for modern telecommunications and voice interface systems, especially in noisy acoustic environments. This study is important due to the limited investigation of contemporary hybrid neural network architectures for VAD in low-resource languages such as Kazakh, particularly across a wide range of SNR levels, including extreme values below -10 dB. The central research question is which modern hybrid neural network architecture offers the best balance between accuracy and computational efficiency for speech detection in the Kazakh language under severe noise conditions. This study developed and tested five architectures, CNN+BiGRU, CNN+GRU, CNN+LSTM, CNN+BiLSTM, and CNN+TDNN, based on the KSC2 corpus, augmented with synthetic noise across an SNR range from -18 dB to +30 dB, with separate analyses at fixed levels of 10 dB and -10 dB. MFCC features were used as input, and training/testing was performed using noise samples from the ESC-50 dataset. Experimental results demonstrated that the CNN+BiGRU, CNN+GRU, and CNN+LSTM architectures achieved the highest F1-score (99.6%) and maintained robustness at SNR levels above -12 dB, whereas CNN+TDNN provided comparable quality with minimal computational complexity and the shortest training time (164 s). The analysis under fixed SNR levels revealed the limited generalization capabilities of the models when trained on a single noise level, highlighting the necessity of incorporating a wide SNR range in training. In conclusion, the hybrid architectures CNN+BiGRU and CNN+TDNN are recommended for deployment in VAD systems for the Kazakh language in highly noisy environments.

Keywords-Voice Activity Detection (VAD); Low Signal-to-Noise Ratio (SNR); Convolutional Neural Network (CNN); Recurrent Neural Network (RNN); BiGRU; BiLSTM

I. INTRODUCTION

Voice Activity Detection (VAD) plays a crucial role in modern speech processing systems, as it is employed as a preliminary stage in tasks such as Automatic Speech Recognition (ASR), speech signal encoding and transmission, speaker diarization, speaker identification, and telecommunication and multimedia systems. The performance of VAD modules directly affects the efficiency of subsequent stages in speech signal processing, including recognition accuracy, system response time, transmission channel throughput, and overall computational performance. VAD can be formalized as a binary classification task [1] aimed at separating incoming audio signals into speech and nonspeech segments. The accuracy of this segmentation determines both the reliability of downstream processing and the overall robustness of the speech system to noise and interference.

One of the key challenges faced by VAD system developers is ensuring reliable speech detection under low Signal-to-Noise Ratio (SNR) conditions. In real-world scenarios, such as public transportation, industrial environments, open spaces, or multi-speaker interactions, speech signals are accompanied by background noise, which significantly complicates the separation of speech segments from non-speech ones. At SNR levels below 0 dB, traditional VAD methods exhibit a sharp decline in accuracy, underscoring the need for new approaches that are robust to severe noise distortions.

Traditional VAD algorithms rely on simple statistical features of the signal, such as energy, Zero-Crossing Rate (ZCR), spectral entropy, Linear Predictive Coding (LPC), and spectral subtraction. These methods are characterized by low computational complexity and high processing speed, making them suitable for real-time applications. However, their effectiveness significantly deteriorates under degraded conditions, particularly at SNR levels below 5 dB [2, 3].

More advanced approaches include algorithms embedded in industrial codecs, such as G.729B, AMR, and WebRTC. These methods utilize spectral energy, threshold-based segmentation, and heuristic rules, demonstrating acceptable performance only under moderate noise conditions (SNR > 10 dB). At SNR levels below 0 dB, such systems tend to produce a significant number of false positives and missed speech segments (i.e., false negatives) [4]. To enhance the noise robustness of VAD systems, various noise suppression methods have been proposed, including Spectral entropy Minimization (SMPR), spectral subtraction, Discrete Wavelet Transform (DWT), and their combinations with machine-learning algorithms. In [5], a relatively simple VAD algorithm was based on Short-Term Energy (STE), periodicity, and spectral Flatness (SF), achieving high accuracy only under moderate noise conditions (SNR $\geq$ 0 dB) and rapidly losing its effectiveness in more complex acoustic environments. Further progress in this direction was demonstrated in [6], developing a lightweight and reliable algorithm, mVAD, using hybrid features: MFCC, LPC, and Normalized Spectral Centroid (NSSC). This method showed significant improvements over traditional VAD algorithms in terms of accuracy and latency, maintaining operational reliability even at SNR levels below 0 dB.

In addition to these approaches, a statistical model-based VAD method in [7] aimed to reduce the frequency of false Likelihood Ratio Test (LRT) rejections under noisy conditions. Although this method achieved a reduction of 15.8% in error rate compared to conventional VAD, its advantages were observed only within a positive SNR range of 0-20 dB. Additionally, the study in [8] contributed to the field by developing a VAD method based on spectral entropy computation using filter bank outputs. This approach significantly reduced the computational complexity compared to Fast Fourier Transform (FFT)-based methods and achieved higher accuracy at SNR levels below 0 dB. However, the general applicability of this approach is limited by the narrow range of testing conditions.

In parallel with the development of traditional and hybrid VAD algorithms, researchers have begun to integrate more complex method combinations to improve robustness under challenging acoustic conditions. For instance, in [9, 10] it was shown that hybrid architectures such as SS+DWT+SVM can achieve high performance metrics (e.g., PESQ and STOI) even at SNR levels of -10 dB. However, these approaches require meticulous manual feature tuning and are insufficiently robust in dynamically changing noise environments. In [11], a dynamic noise filtering method, D-FBSS, employed different Deep Neural Network (DNN) models for each type of background noise. This approach improves the detection accuracy in noisy environments but requires a large amount of pre-labeled data and substantial computational resources.

In general, traditional VAD methods demonstrate good efficiency under steady-state noise conditions, but their performance decreases dramatically as the noise characteristics change over time. For example, in real acoustic environments, such as restaurants or outdoor settings, the spectral characteristics of noise can change continuously, leading to a decrease in VAD accuracy [12]. This problem highlights the importance of using VAD models in the learning process, not only artificial noise (for example, white noise), but also acoustically realistic noises that reflect the conditions of the everyday environment. Combining such types of noise makes it possible to enhance the generalizability of the models and improve their robustness to various types of background interference.

Combining such types of noise makes it possible to enhance the generalizability of the models and improve their robustness to various types of background interference. Recent studies have shown that such approaches outperform traditional algorithms in terms of both accuracy and generalization ability under noisy conditions [13, 14]. Convolutional Neural Networks (CNNs) have proven particularly effective in extracting time-frequency features that are robust to noise and speaker variability [15]. Recurrent neural networks, such as LSTM and GRU, can capture temporal dependencies, which are critically important for VAD, as the current speech frame depends on its surrounding context [16]. The use of bidirectional networks, such as BiLSTM and BiGRU, improves recognition completeness by considering both past and future contexts; however, such models have latency limitations and are not always suitable for real-time systems [17]. Time Delay

Neural Networks (TDNN), originally developed for ASR tasks, deserve particular attention in this regard, as they have demonstrated the ability to effectively model temporal dependencies with low latency and have been adapted for use in VAD, showing strong robustness under noisy conditions [18].

Recent studies have demonstrated further advances in neural network-based VAD models aimed at improving their accuracy in real-world acoustic environments. For instance, in [19], an efficient real-time neural network-based VAD algorithm was able to deliver state-of-the-art performance in complex real-world recordings. This study demonstrated that the use of segmental SNR yields more stable results than methods based on clean-speech characteristics. However, the experiments were conducted only at positive SNR levels (5 and 10 dB), which limits the evaluation of the effectiveness under extreme noise conditions. In [20], a hybrid feature extraction method combined MFCC and DWT with denoising through a median filter in the wavelet domain, implemented on a Raspberry Pi 3. This method showed very high recognition rates in clean environments and satisfactory results even at various noisy SNRs. Although it does not explicitly focus on VAD, its robustness in feature extraction under noisy conditions is relevant because feature quality strongly impacts VAD performance.

In [21], a speech preprocessing system was based on a Deep Neural Network (DNN) to improve VAD accuracy under noisy conditions. The test results on the NOIZEUS database [22] showed that this approach outperformed VQ-VAD; however, the experiments were also limited to an SNR range of -5 to 10 dB. In [23], the DWT-CNN-MCSE architecture was proposed, which demonstrated a significant improvement in speech recognition accuracy at an SNR of 10 dB. Such architectures enable effective extraction of both spatial and temporal features; however, most existing studies remain limited since training is typically conducted at fixed noise levels (-5, 0, and 5 dB), more extreme SNR conditions below -10 dB are rarely considered, the focus is predominantly on English-language corpora, and computational complexity is scarcely evaluated [24].

Table I summarizes the most representative VAD methods from classical statistical models to recent deep learning approaches, with emphasis on performance under low-SNR conditions. Classical methods, such as GMM-HMM and WebRTC VAD, suffer from high missed detection rates under low SNR, while recent transformer-based models achieve high accuracy at the cost of excessive computational complexity. Thus, despite progress in the field of VAD, the problem of robust speech detection at SNR levels as low as -20 dB remains unsolved. This problem is particularly relevant for low-resource languages such as Kazakh, which is characterized by agglutinative morphology, vowel harmony, and variable intonation patterns, which reduce the effectiveness of models trained on English-language data.

This study aimed to address this gap. Its objective was to design and comparatively evaluate five hybrid neural network VAD architectures for the Kazakh language across an SNR range from -18 dB to +30 dB, including analysis at fixed levels of 10 dB and -10 dB, followed by an assessment of classification accuracy and computational efficiency. This study utilizes realistically noise-contaminated data from the Kazakh Speech Corpus (KSC2) [34] and noise samples from ESC-50 [35], and conducts a comparative analysis based on Accuracy, Recall, Precision, F1-score, and computational complexity metrics (number of parameters and training time)

TABLE I. COMPARISON OF CLASSIC AND MODERN VAD METHODS UNDER LOW-SNR CONDITIONS

Year	Method	Data/Noise	Conditions, SNR	Ref
2007	Energy-based + hangover	Algorithmic baseline	Degrades at very low SNR	[25]
1999	HMM hangover	Algorithmic baseline	Designed for noisy conditions	[26]
1996/ 2001	ITU-T G.729 Annex B VAD (standard)	Telephony codec context	Real-time/tight latency	[27]
2006/ 2022	3GPP TS 26.094 AMR/AMR-WB VAD (standard)	Mobile (DTX)	Real-time/tight latency	[28]
2021	CNN-BiLSTM	AVA-Speech [14]	Movie audio;	[17]
2024	Transformer-VAD	AVA-Speech [14]	Real-world; multilingual; challenging noise	[29]
2024	PVAD comparative	Real-world PVAD setups	Seen/unseen speakers and noise	[30]
2023	TS-VAD with Transformers	Diarization context	Streaming/causal variants	[31]
2022	Tr-VAD	Benchmark speech/ noise corpora	Low-SNR robustness focus	[32]
2025	SincQDR-VAD	Multiple benchmarks	Low-SNR efficiency-oriented	[33]

II. PROPOSED METHOD

The KSC2 dataset, developed by the Institute of Smart Systems at Nazarbayev University [34], was used as the primary dataset for this study. The corpus contains recordings from 169 speakers, each producing 75 different utterances. For this study, recordings from the first 30 speakers were selected, resulting in 2,250 original audio files. To enhance the robustness of the models to various acoustic interference conditions, all recordings were augmented with white Gaussian noise at SNR levels from -18 dB to 30 dB in 3 dB increments. Consequently, each audio file was duplicated into 13 noise-augmented variants, with each copy stored in a separate folder. This approach enabled simulation of a wide range of real-world acoustic scenarios, including extremely low SNR conditions.

During the preprocessing stage, an automatic segmentation algorithm was implemented to divide audio recordings at the word level using wrd annotation files containing timestamps for the start and end of each word. As a result, a large-scale dataset was generated in which each segment corresponded to an individual word form. A total of 1,509,284 segments were generated for all SNR levels. The data were split into training and test sets in an 80:20 ratio by speaker, ensuring that the test set included only speakers who were not present in the training set. This approach provides a rigorous evaluation of the models' ability to generalize to unseen voices. Consequently, 2,493,440 training segments and 623,360 test segments were formed, with the test set accounting for approximately 20% of the total dataset.

The open ESC-50 dataset [35] was used for noise augmentation. It contains 2,000 labeled audio recordings of environmental sounds divided into 50 categories. Four noise categories were selected for the experiments: animal sounds, natural noises, household noises, and urban noises. White noise was used as a supplementary noise source to ensure controlled experimental conditions. An equal number of files was used for each type of noise, with a uniform distribution across the training and test sets. All noise recordings were resampled at a sampling rate of 16 kHz and scaled according to their Root Mean Square (RMS) amplitude to achieve the desired SNR level when mixed with speech segments.

To represent the audio fragments in a form suitable for machine learning, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each file. The feature extraction parameters included a sampling rate of 16 kHz, analysis window length of 20 ms (320 samples), frame shift of 5 ms (80 samples), 25 Mel filter banks, and 25 MFCC coefficients. However, for further analysis, only the 2nd to 25th coefficients were used; the first coefficient, which represents the overall signal energy, was excluded to reduce the risk of overfitting the model. Consequently, each speech segment was transformed into a feature matrix of size 24×24 (coefficients × time frames). All features were normalized to the range [0–1]. The selection of these parameters aligns with established practices in the field of speech processing and provides adequate time-frequency resolution for analyzing short segments under noisy acoustic conditions.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 reshape (Reshape)           (None, 24, 24, 1)         0

 conv2d (Conv2D)             (None, 24, 24, 16)        160

 max_pooling2d (MaxPooling2D  (None, 12, 12, 16)       0
 )

 conv2d_1 (Conv2D)           (None, 12, 12, 16)        2320

 max_pooling2d_1 (MaxPooling  (None, 6, 6, 16)         0
 2D)

 reshape_1 (Reshape)         (None, 36, 16)            0

 bidirectional (Bidirectiona  (None, 36, 32)           3264
 l)

 bidirectional_1 (Bidirectio  (None, 32)               4800
 nal)

 dense (Dense)               (None, 16)                528

 dense_1 (Dense)             (None, 2)                 34

=================================================================
Total params: 11,106
Trainable params: 11,106
Non-trainable params: 0
```

Fig. 1. Architecture of the CNN+BiGRU model

To address the VAD task, five hybrid architectures combining convolutional and recurrent neural networks were developed and tested: CNN+BiGRU, CNN+GRU, CNN+LSTM, CNN+BiLSTM, and CNN+TDNN. The architecture of the CNN+BiGRU model, which was selected as a baseline, is presented in detail in Figure 1. The CNN+BiGRU model was selected as the baseline, as it combines the ability to extract spatial features using convolutional layers with the effective processing of temporal dependencies in a bidirectional format, allowing the model to consider context from both forward and backward directions.

The remaining architectures differed in the configuration of the recurrent layers while maintaining the overall structure: the CNN+GRU model used unidirectional GRU layers instead of bidirectional ones, CNN+LSTM employed unidirectional LSTM layers, CNN+BiLSTM incorporated bidirectional LSTM layers, and CNN+TDNN utilized a TDNN block after the initial convolutional layers, based on one-dimensional convolutions (Conv1D) followed by a global average pooling layer (GlobalAveragePooling1D).).). For all models, the ReLU activation function was used in the internal layers, and softmax was applied in the output layer. The class labels "speech" and "noise" were encoded in a one-hot format to ensure compatibility with the categorical crossentropy loss function. Training was conducted over 10 epochs using the Adam optimizer with a fixed learning rate of 0.001 and a batch size of 1,024 samples.

All experiments were conducted on a high-performance computing platform equipped with an Intel Core i9-14900KF processor (24 threads, 3.20 GHz), 128 GB DDR5 RAM, and an NVIDIA GeForce RTX 4090 GPU with 24 GB video memory, running on Windows 11 Pro. The software implementation was carried out in Python 3.9 using the following libraries: Librosa for audio feature extraction, NumPy and Pandas for data processing, TensorFlow/Keras for building and training neural network models, and Scikit-learn for metric computation and result validation.

To quantitatively assess the performance of the models, the following evaluation metrics were used:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} - \text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (4)$$

The SNR used for augmentation and evaluation was defined as:

$$\text{SNR(dB)} = 10\log_{10}\left(\frac{P_{speech}}{P_{noise}}\right) \quad (5)$$

The evaluation was conducted on the average results across all SNR levels and on specific values, including fixed levels of 10 dB and -10 dB, and the range of -18 dB to +30 dB. During training, the loss and accuracy values were recorded after each epoch for both training and validation sets to monitor convergence and control overfitting. The training dynamics

plots demonstrated a consistent decrease in error and an increase in accuracy in both the training and validation sets, indicating a stable convergence of the models. During dataset construction, class balance between "speech" and "non-speech" segments was maintained in both the training and test sets.

III. RESULTS

This section presents the results of a comparative analysis of the five tested neural network VAD architectures under various SNR conditions, including fixed values of 10 dB and -10 dB, as well as a broad range from -18 dB to +30 dB. The evaluation focuses on key classification quality metrics, computational complexity, and the robustness of the models in noisy environments.

A. Training Characteristics of the Models

Table II presents the training characteristics of the models. All architectures were trained under the same conditions for 10 epochs with a batch size of 10. This table includes the accuracy and loss values for the training and test sets after training is completed, as well as the number of trainable parameters for each architecture. The loss values across all models ranged from 9.8% to 10.5%, while the accuracy on both datasets remained at 96% for all of them. All models demonstrated a consistent decrease in the loss function throughout all training epochs without signs of sharp fluctuations or divergence between the training and validation sets, which is confirmed by the close values of loss and accuracy at the end of training.

TABLE II. MODEL TRAINING RESULTS: ACCURACY, LOSS, AND NUMBER OF PARAMETERS

Model	Epochs	Accuracy (train/test)	Loss (train/test)	Number of parameters
CNN+BiGRU	10	96%/96%	9.8%/9.9%	11,106
CNN+BiLSTM	10	96%/96%	9.8%/9.8%	13,538
CNN+GRU	10	96%/96%	10.2%/10.1%	6,050
CNN+LSTM	10	96%/96%	10.1%/10.3%	7,010
CNN+TDNN	10	96%/96%	10.4%/10.5%	5,650

The number of parameters ranged from 5,650 (CNN+TDNN) to 13,538 (CNN+BiLSTM), reflecting the varying degrees of structural complexity among the tested architectures. The CNN+BiGRU and CNN+BiLSTM models have the highest number of parameters, while CNN+TDNN and CNN+GRU exhibit lower architectural complexity, which may potentially affect computational costs during training and practical deployment.

B. Analysis of Training Results at Fixed SNR Levels

A series of experiments was performed. In the initial stage, the selected neural network models were trained at fixed SNR levels, such as -10, 0, 10, and 20 dB. After training, each model was tested across the full range of SNR values. Figure 2 presents the results of training the models at an SNR of 10 dB and testing them across the entire range. All investigated models demonstrated high accuracy at positive SNR values, confirming their effectiveness under low-noise interference. However, as the SNR level decreased, particularly in the negative range (below 0 dB), a significant drop in the classification accuracy was observed.

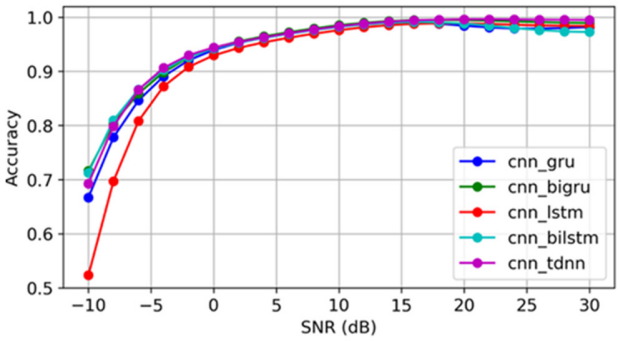

Fig. 2. Testing the models trained at a fixed SNR level of 10 dB.

Figure 3 shows the results of training the models at a fixed SNR level of -10 dB and their subsequent testing across the entire range of SNR values. It can be observed that when training the models at an SNR level of -10 dB, the models demonstrated high accuracy, close to the training value, indicating their ability to effectively reproduce the conditions under which they were trained. However, as the SNR level increases, there is a sharp decline in accuracy, which may be attributed to the models perceiving the new acoustic conditions as deviations from the training distribution and interpreting them as outliers.

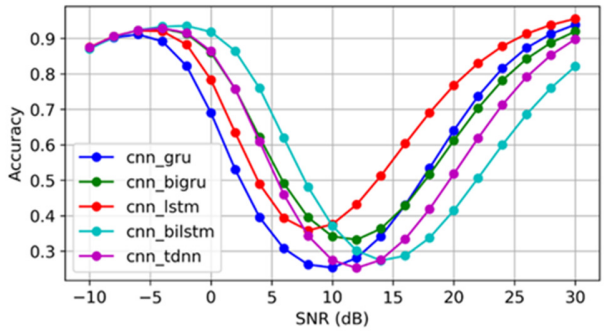

Fig. 3. Testing the models trained at a fixed SNR level of -10 dB.

This indicates a low generalization capability of the models when they are trained only within a narrow range of noise conditions. This result highlights the need for a training dataset that includes a wide range of SNR levels to enhance the robustness of the models to the acoustic variability of real-world speech signals. Consequently, in the next stage of the experiment, the models were trained across the entire range of SNR values. This strategy made it possible to reveal the impact of the training SNR range on the model performance and to perform a comparison with the results obtained from the training on the extended dataset.

C. Comparison of Models in Terms of Accuracy Across the Entire SNR Range

Figure 4 shows the classification accuracy of all models trained on SNRs in the range of -18 dB to +30 dB. This provides insight into the robustness of classification performance under varying noise conditions.

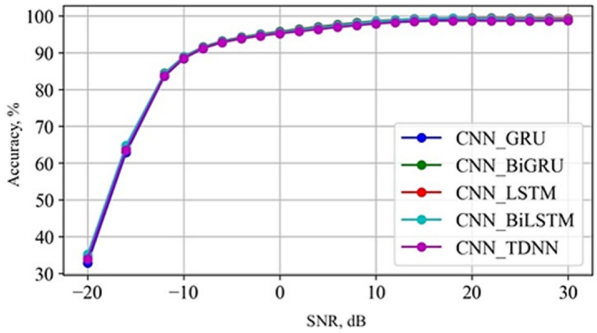

Fig. 4. Accuracy of the models trained in SNR of -18 to +30 dB.

All models demonstrated a general increase in classification accuracy with increasing SNR levels. For SNR values above 0 dB, the accuracy exceeded 95% across all models. In the negative SNR range (from -20 dB to 0 dB), differences in performance become evident: the CNN+BiGRU and CNN+LSTM architectures maintain higher accuracy across the entire noise level spectrum, whereas the CNN+TDNN and CNN+GRU models show a noticeable drop in accuracy under extremely low SNR conditions (below -10 dB).

Compared with the results presented in Figure 2, where the training was performed only at a fixed SNR level of 10 dB, the models trained across a wide range of SNR values exhibited significantly higher accuracy at low SNR levels. For instance, at an SNR of -10 dB, the accuracy of the CNN+BiGRU model increased from 73% to 84%. This comparative trend in accuracy illustrates the differences in the ability of various architectures to maintain correct classification as the acoustic conditions deteriorate.

D. Comparison of Architecture Efficiency

All models were trained and tested on a unified dataset to reflect various levels of SNR, including fixed values of 10 dB and -10 dB, as well as a wide range from -18 dB to +30 dB in 3 dB increments. This approach enabled for an objective and comparable evaluation of model performance under different acoustic conditions. The classification quality was assessed using standard binary classification metrics, F1-score, accuracy, precision, and recall, which are commonly used in VAD tasks and provide a comprehensive characterization of the model behavior in noisy environments. Table III presents the average results for each architecture tested across all SNR levels ranging from -18 to +30 dB. These data provide insights into both the overall accuracy of the models and their ability to correctly classify speech and nonspeech segments under background noise conditions.

TABLE III. COMPARISON OF DIFFERENT NEURAL NETWORK ARCHITECTURES UNDER VARYING SNR LEVELS

Model	F1 %	Precision %	Accuracy %	Recall %
CNN+BiGRU	99.6	99.4	99.6	99.8
CNN+BiLSTM	99.5	99.2	99.6	99.8
CNN+LSTM	99.6	99.5	99.7	99.8
CNN+TDNN	99.3	98.7	99.3	99.9
CNN+GRU	99.6	99.3	99.6	99.9

Among the models tested, the highest F1-scores (99.6%) were obtained by the CNN+BiGRU, CNN+LSTM, and CNN+GRU models, indicating a strong consistency in accurately classifying speech and nonspeech segments across the full range of noise levels. The highest accuracy (99.7%) was observed for the CNN+LSTM model, demonstrating its superior overall prediction accuracy averaged across all testing conditions. Simultaneously, the CNN+TDNN model demonstrated the lowest average accuracy (99.3%) among the tested architectures; however, it achieved the highest recall (99.9%), indicating its high sensitivity in detecting speech events, including scenarios with low SNR levels. These results highlight the trade-offs between precision and recall across different architectures, which may be crucial for their application in practical systems.

E. Model Training Time

In addition to quality metrics and the number of parameters, training time is an important factor when comparing models. Figure 5 presents the training times for each model under identical experimental conditions. The training was conducted for 10 epochs with a batch size of 1024 on the same computing platform. The longest training times were recorded for the CNN+BiLSTM (558 s) and CNN+BiGRU (525 s) models, due to their more complex architectures and larger number of trainable parameters. The CNN+GRU and CNN+LSTM models required 300 s and 318 s for training, respectively. The CNN+TDNN model demonstrated the shortest training time (164 s), reflecting its compact architecture and lower computational load compared to the other tested models. All training time measurements were performed under identical experimental settings to ensure the comparability of the obtained values across different architectures.

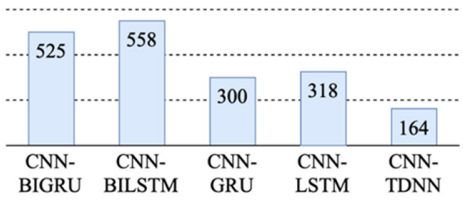

Fig. 5. Training time of the models on the training set (in seconds)

F. F1-Score Dependence on SNR

Figure 6 shows the dependence of the F1-score on the SNR for the CNN+BiGRU model. This characteristic reflects the variation in classification quality under changing acoustic conditions and provides a quantitative assessment of the model's sensitivity to increasing noise levels. At an SNR of -20 dB, the F1-score was 46%, indicating a significant drop in classification accuracy under extreme noise conditions. As the SNR increased to -12 dB, the F1-score reached 91%, demonstrating a substantial recovery in the classification performance, even at negative SNR levels. At an SNR of -20 dB, the F1-score was 46%, indicating a significant drop in classification accuracy under extreme noise conditions. As the SNR increased to -12 dB, the F1-score reached 91%, demonstrating a substantial recovery in classification

performance, even at negative SNR levels. In the range of -18 to -4 dB, the F1-score increased more sharply, reflecting the highest sensitivity of the model under heavily noisy conditions. This trend characterizes the operational boundaries of the model and defines the range of SNR values at which the target classification performance was achieved.

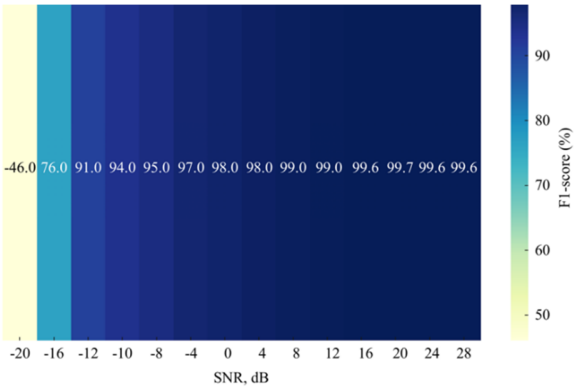

Fig. 6. F1-score dependency on SNR for the CNN+BiGRU model.

G. Confusion Matrices at Low SNR Levels

To analyze the nature of the classification errors, an SNR level of -6 dB was selected, as all the models tested remained operational at this level, and the differences in the distribution of errors between the architectures became more apparent. This allows for a quantitative comparison of False Positives (FP) and False Negatives (FN) under relatively challenging, yet not extreme, acoustic conditions. Figure 7 presents the normalized confusion matrices for the CNN+BiGRU, CNN+BiLSTM, CNN+GRU, and CNN+LSTM models at an SNR of -6 dB. Each matrix illustrates the ratio of correctly and incorrectly classified segments for the "speech" and "noise" classes.

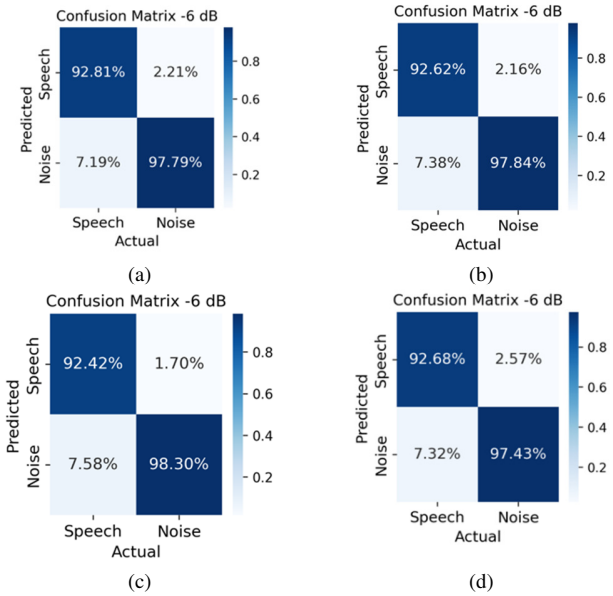

Fig. 7. Confusion matrices at low SNR level: (a) CNN+BiGRU, (b) CNN+BiLSTM, (c) CNN+GRU, (d) CNN+LSTM.

For the CNN+BiGRU model, the classification accuracies for speech and noise segments were 92.81% and 97.79%, respectively, while the proportion of false positives was 2.21%, and that of false negatives was 7.19%. For the CNN+BiLSTM model, the recognition accuracy for speech segments was 92.62%, and for noise segments, 97.84%, while the false positive rate was 2.16% and the false negative rate was 7.38%. The CNN+GRU model achieved an accuracy of 92.42% for speech segments and 98.30% for noise segments, while the false-positive rate was 1.70% and the false-negative rate was 7.58%. For the CNN+LSTM model, the recognition accuracy for speech segments was 92.68% and for noise segments was 97.43%, while the false-positive rate was 2.57% and the false-negative rate was 7.32%. These results allow for a comparative assessment of the behavior of the tested architectures under high-noise conditions and provide insight into the balance between the accuracy of recognizing speech and non-speech segments at a fixed level of acoustic interference.

IV. DISCUSSION

In [36], a comparative analysis was conducted on the effectiveness of convolutional and recurrent neural networks for VAD based on clean speech signals. In [37], this approach was extended by training the models at a fixed SNR of 20 dB. This work represents a continuation of this research and focuses on evaluating the effectiveness of hybrid architectures that combine convolutional and recurrent layers under low SNR conditions, enabling an assessment of their robustness to acoustic interference and variability in the noise environment.

The experimental results confirm the high effectiveness of the hybrid neural network architectures tested in the VAD task under varying levels of acoustic noise. All five models demonstrated high performance metrics across different SNR levels, indicating their potential applicability in real-world scenarios. The highest F1-score (99.6%) was achieved by the CNN+BiGRU, CNN+GRU, and CNN+LSTM architectures, indicating their balanced ability to correctly classify speech segments while maintaining a low false-positive rate. The CNN+LSTM model achieved the highest average accuracy (99.7%), demonstrating its high prediction precision in noisy acoustic environments. The CNN+GRU and CNN+TDNN architectures achieved the highest recall values (99.9%), which is particularly important in applications where minimizing missed speech events is a priority.

Particular attention should be paid to the CNN+TDNN model, which, despite having the smallest number of parameters (5,650), demonstrated a classification performance comparable to that of more complex architectures. This characteristic makes CNN+TDNN a promising option for deployment in resource-constrained devices, where a balance between accuracy and computational load must be maintained.

A separate analysis was conducted for scenarios in which training was performed at fixed SNR levels (10 dB and -10 dB). The results showed that such models demonstrated high accuracy near the training SNR level but experienced rapid degradation in performance when the conditions deviated from the training distribution. For example, models trained at an SNR of 10 dB exhibited a sharp drop in accuracy at negative

SNR values, whereas models trained at an SNR of -10 dB failed to maintain high accuracy at positive SNR levels. This highlights the limitations of training solely at fixed noise levels and emphasizes the need to use training datasets that cover a wide range of conditions to develop robust and generalizable models.

Analyzing the dependence of the F1-score on SNR for the CNN+BiGRU model showed that an F1-score above 90% was achieved at SNR of -12 dB, and at SNR ≥ 0 dB, the metric stabilized at 99% or higher. This confirms the ability of CNN+BiGRU to maintain a high classification quality even under severe noise conditions. All tested models show a consistent recovery of accuracy with increasing SNR: when SNR ≥ -10 dB, accuracy drops sharply, but at SNR of -8 dB, it already exceeds 90%, and at SNR greater than 0 dB, it recovers to values between 95% and 99%. The CNN+BiGRU and CNN+LSTM architectures demonstrated the smoothest improvement in performance as the SNR increased, confirming their adaptability and strong generalization capability under diverse acoustic conditions.

The confusion matrices at a fixed SNR level of -6 dB allow the assessment of the error patterns of the tested models under heavily noisy conditions. All architectures exhibited higher classification accuracy for noise segments (~98%) than for speech segments (ranging from 90.6% to 92.8%). This is a typical feature of VAD tasks under negative SNR conditions, where the speech signals are partially masked by noise. In this context, the CNN+BiGRU and CNN+LSTM models exhibited the lowest false-positive (FP) and false-negative (FN) error rates, indicating their reliability in detecting voice activity in challenging acoustic scenarios.

A comparative analysis of training time and number of parameters revealed an expected correlation between architectural complexity and computational cost. Architectures with bidirectional recurrent layers (CNN+BiGRU and CNN+BiLSTM) required the longest training times (525 and 558 s, respectively) and contained the highest number of parameters (11,106 and 13,538), indicating their more resource-intensive structure. Simultaneously, the CNN+GRU, CNN+LSTM, and especially the CNN+TDNN models demonstrated significantly lower training times (ranging from 300 to 164 s) while maintaining high levels of accuracy and F1-score, making them preferable for deployment in applications with limited computational resources.

The experimental results confirm the practical applicability of the models developed for automatic voice activity detection tasks in noisy conditions, including scenarios with negative SNR. The CNN+BiGRU and CNN+TDNN architectures are of particular interest in terms of the balance between classification accuracy and computational efficiency, making them suitable candidates for implementation in resource-constrained devices, such as mobile and embedded systems.

However, this study has several limitations that should be taken into account:

- The use of synthetically noised data (by adding noise to clean recordings) may not fully reflect the complexity of real-world acoustic scenes with their temporal and spectral characteristics.
- The experiment did not consider the real-time operation of the models, including processing latency and robustness during continuous signal input.
- The models were trained and tested on relatively short audio segments. Processing long continuous audio streams may require additional measures to ensure stability and classification quality.

V. CONCLUSIONS

This study presents an experimental investigation of the effectiveness of five hybrid neural network architectures, CNN+BiGRU, CNN+GRU, CNN+LSTM, CNN+BiLSTM, and CNN+TDNN, for solving the problem of VAD at varying levels of acoustic noise. The analysis was conducted using the KSC2, augmented with synthetic noise generated from the ESC-50 dataset and white noise at SNR levels ranging from -18 to +18 dB. MFCCs were used as input features, enabling the construction of compact and informative representations of audio signals for input into the models.

The experimental results showed that all architectures demonstrated high classification performance across the entire range of testing conditions (accuracy > 99.3%, F1-score > 99.3%). The CNN+BiGRU architecture achieved the best balance of precision, recall, and classification robustness, reaching an F1-score of 99.6% and maintaining a stable performance at SNR levels above -12 dB. The CNN+TDNN model demonstrated a solid performance with minimal computational complexity and the shortest training time, making it attractive for use in applications with limited computational resources. The confusion matrix analysis confirmed the ability of the tested architectures to effectively distinguish between speech and noise segments, even under severe acoustic interference (e.g., at an SNR of –6 dB). Moreover, the analysis of models trained at fixed SNR levels (10 dB and -10 dB) revealed their limited generalization capability beyond the training conditions, highlighting the importance of a training dataset with a wide range of noise levels to enhance the robustness of the model in real-world acoustic scenarios. The proposed architectures, particularly CNN+BiGRU and CNN+TDNN, demonstrate strong potential for deployment in practical voice activity detection systems, such as voice assistants, telecommunication platforms, and monitoring devices operating in noisy environments.

Future work is planned to expand this research by conducting real-time model testing, exploring the adaptability of the model to non-standard noise conditions, and developing approaches for online model updating to ensure robustness in dynamically changing acoustic environments.

ACKNOWLEDGMENT

This research is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP22684173) "Development of a highly efficient neural network method for detecting voice activity at a low signal-to-noise ratio".

REFERENCES

[1] R. M. Patil and C. M. Patil, "Unveiling the State-of-the-Art: A Comprehensive Survey on Voice Activity Detection Techniques," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, MYSORE, India, Jul. 2024, pp. 1–5, https://doi.org/10.1109/APCIT62007.2024.10673721.

[2] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, vol. 50, no. 6, pp. 453–466, June 2008, https://doi.org/10.1016/j.specom.2008.01.003.

[3] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, Sept. 2007, https://doi.org/10.1121/1.2766778.

[4] M. A. Hasby, A. G. Putrada, and F. Dawani, "The Quality Comparison of WebRTC and SIP Audio and Video Communications with PSNR," *Indonesian Journal on Computing*, vol. 6, no. 1, pp. 73–84, Apr. 2021.

[5] R. Çolak and R. Akdeniz, "A Novel Voice Activity Detection for Multi-Channel Noise Reduction," *IEEE Access*, vol. 9, pp. 91017–91026, 2021, https://doi.org/10.1109/ACCESS.2021.3086364.

[6] Z. Zhu, L. Zhang, K. Pei, and S. Chen, "A robust and lightweight voice activity detection algorithm for speech enhancement at low signal-to-noise ratio," *Digital Signal Processing*, vol. 141, Sept. 2023, Art. no. 104151, https://doi.org/10.1016/j.dsp.2023.104151.

[7] S. M. Kim, "Auditory Device Voice Activity Detection Based on Statistical Likelihood-Ratio Order Statistics," *Applied Sciences*, vol. 10, no. 15, Jan. 2020, Art. no. 5026, https://doi.org/10.3390/app10155026.

[8] F. Liu and A. Demosthenous, "A Computation Efficient Voice Activity Detector for Low Signal-to-Noise Ratio in Hearing Aids," in *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Dec. 2021, pp. 524–528, https://doi.org/10.1109/MWSCAS47672.2021.9531915.

[9] Y. Iqbal *et al.*, "A Hybrid Speech Enhancement Technique Based on Discrete Wavelet Transform and Spectral Subtraction," *IEEE Access*, vol. 13, pp. 39765–39781, 2025, https://doi.org/10.1109/ACCESS.2025.3546434.

[10] B. G. Nagaraja, G. T. Yadava, P. Kabballi, and C. M. Patil, "VAD system under uncontrolled environment: A solution for strengthening the noise robustness using MMSE-SPZC," *International Journal of Speech Technology*, vol. 27, no. 2, pp. 309–317, Jun. 2024, https://doi.org/10.1007/s10772-024-10104-w.

[11] M. Aliouat and M. Djendi, "A new deep learning forward BSS (D-FBSS) algorithm for acoustic noise reduction and speech enhancement," *Applied Acoustics*, vol. 230, Feb. 2025, Art. no. 110413, https://doi.org/10.1016/j.apacoust.2024.110413.

[12] U. Shrawankar and V. Thakare, "Voice Activity Detector and Noise Trackers for Speech Recognition System in Noisy Environment," *International Journal of Advancements in Computing Technology*, vol. 2, no. 4, pp. 107–114, Oct. 2010, https://doi.org/10.4156/ijact.vol2.issue4.11.

[13] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Feb. 2013, pp. 7378–7382, https://doi.org/10.1109/ICASSP.2013.6639096.

[14] X. Miao, I. McLoughlin, and Y. Yan, "A New Time-Frequency Attention Mechanism for TDNN and CNN-LSTM-TDNN, with Application to Language Identification," in *Interspeech 2019*, Sep. 2019, pp. 4080–4084, https://doi.org/10.21437/Interspeech.2019-1256.

[15] Y. Tan and X. Ding, "Heterogeneous Convolutional Recurrent Neural Network with Attention Mechanism and Feature Aggregation for Voice Activity Detection," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024, https://doi.org/10.1561/116.00000158.

[16] I. Han, C. N. Om, and U. I. Kim, "A gated recurrent unit based robust voice activity detector," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 41939–41949, Apr. 2024, https://doi.org/10.1007/s11042-023-17123-w.

[17] N. Wilkinson and T. Niesler, "A Hybrid CNN-BiLSTM Voice Activity Detector," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 6803–6807, https://doi.org/10.1109/ICASSP39728.2021.9415081.

[18] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Interspeech 2017*, Aug. 2017, pp. 999–1003, https://doi.org/10.21437/Interspeech.2017-620.

[19] S. Braun and I. Tashev, "On training targets for noise-robust voice activity detection," in *2021 29th European Signal Processing Conference (EUSIPCO)*, Dec. 2021, pp. 421–425, https://doi.org/10.23919/EUSIPCO54536.2021.9616082.

[20] A. Mnassri, M. Bennasr, and C. Adnane, "A Robust Feature Extraction Method for Real-Time Speech Recognition System on a Raspberry Pi 3 Board", *Engineering, Technology, & Aplied Science Research*, vol. 9, no. 2, pp. 4066–4070, Apr. 2019.

[21] B. G. Nagaraja and G. T. Yadava, "Enhancing Voice Activity Detection in Noisy Environments Using Deep Neural Networks," *Circuits, Systems, and Signal Processing*, vol. 44, no. 7, pp. 5220–5234, July 2025, https://doi.org/10.1007/s00034-025-03055-3.

[22] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, July 2007, https://doi.org/10.1016/j.specom.2006.12.006.

[23] P. Cherukuru and M. B. Mustafa, "CNN-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (DWT) preprocessing," *PeerJ Computer Science*, vol. 10, Feb. 2024, Art. no. e1901, https://doi.org/10.7717/peerj-cs.1901.

[24] N. K. Singh, and Y. J. Chanu, "Robust Voice Activity Detection Algorithm based on Long Term Dominant Frequency and Spectral Flatness Measure," *International Journal of Image, Graphics and Signal Processing*, vol. 9, no. 8, pp. 50–58, Aug. 2017, https://doi.org/10.5815/ijigsp.2017.08.06.

[25] J. Ramírez, J. M. Górriz, J. C. Segura, "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," in *Robust Speech Recognition and Understanding*, 2007.

[26] J. Sohn, N. S. Kim, W. Sung, "A Statistical Model-Based Voice Activity Detector," *IEEE Signal Processing Letters, 1999*.

[27] "Recommendation G.729: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70." International Telecommunication Union ITU), [Online]. Available: https://www.itu.int/rec/T-REC-G.729-199610-S!AnnB/en.

[28] "Specification # 26.194." Third Generation Partnership Project (3GPP) - European Telecommunication Standards Institute (ETSI), [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1428.

[29] B. Karan, J. Jansen van Vüren, F. de Wet, and T. Niesler, "A Transformer-Based Voice Activity Detector," in *Proceedings Interspeech 2024*, Kos, Greece, 2024, pp. 3819–3823, https://doi.org/10.21437/Interspeech.2024-1019.

[30] S. Kumar *et al.*, "Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness." arXiv, June 12, 2024, https://doi.org/10.48550/arXiv.2406.09443.

[31] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target Speaker Voice Activity Detection with Transformers and Its Integration with End-To-End Neural Diarization," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10095185.

[32] Y. Zhao and B. Champagne, "An Efficient Transformer-Based Model for Voice Activity Detection," in *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, Dec. 2022, pp. 1–6, https://doi.org/10.1109/MLSP55214.2022.9943501.

[33] C. C. Wang, E. L. Yu, J. W. Hung, S. C. Huang, and B. Chen, "SincQDR-VAD: A Noise-Robust Voice Activity Detection Framework Leveraging Learnable Filters and Ranking-Aware Optimization." arXiv, Aug. 28, 2025, https://doi.org/10.48550/arXiv.2508.20885.

[34] S. Mussakhojayeva, Y. Khassanov, and H. A. Varol, "KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus," *Interspeech 2022*, Incheon, Korea, 2022, pp. 1367–1371.

[35] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane, Australia, Oct. 2015, pp. 1015–1018, https://doi.org/10.1145/2733373.2806390.

[36] A. Nurlankyzy *et al.*, "The dependence of the effectiveness of neural networks for recognizing human voice on language," *Eastern-European Journal of Enterprise Technologies*, vol. 1, no. 9 (127), pp. 72–81, Feb. 2024, https://doi.org/10.15587/1729-4061.2024.298687.

[37] B. Medetov *et al.*, "Evaluating the effectiveness of a voice activity detector based on various neural networks," *Eastern-European Journal of Enterprise Technologies*, vol. 133, no. 5, pp. 19-28, Jan. 2025, https://doi.org/10.15587/1729-4061.2025.321659.