Evaluating Noise-Robustness of Convolutional and Recurrent Neural Networks for Baby Cry Recognition

Medhanita Dewi Renanti[1], Agus Buono[2], Karlisa Priandana[3], Sony Hartono Wijaya[4]
Doctoral Study Program of Computer Department, IPB University, Bogor, Indonesia[1]
Software Engineering Technology, College of Vocational Studies IPB University, Bogor, Indonesia[1]
Department of Computer, IPB University, Bogor, Indonesia[2, 3, 4]

***Abstract*—Reliable baby cry recognition plays a crucial role in infant care and monitoring, yet real-world environment poses challenges to system accuracy due to its background noises. This study proposes a novel CNN architecture for baby cry recognition under varying noise conditions, featuring three convolutional layers, a max pooling layer, and 0.5 dropout set, and compares its performance against standard RNN models. The models were trained for 100 epochs with a batch size of 64 and evaluated in both clean and noisy environments. To simulate real-world scenarios, recordings were transformed into audio signals and subjected to varying levels of background noise, particularly at different signal-to-noise ratios (SNRs). Results indicate that both models achieved high accuracy (>89%) in noise-free conditions. However, the proposed CNN maintained higher precision (93%) and overall accuracy (91%) than the RNN under 10dB noise, demonstrating its superior noise robustness for baby cry recognition. This improvement is attributed to the CNN's capacity to capture spatial features in audio signals, making it susceptible to noise disruptions. These findings contribute to the development of more reliable and robust baby cry recognition systems.**

Keywords—Baby cry recognition; deep learning; gated recurrent unit; long short-term memory; noise robustness; signal-to-noise ratio

I. INTRODUCTION

Deep learning has firmly established itself as a powerful tool for various tasks, including classification, detection, and noise mitigation. Its ability to improve accuracy and shorten processing duration, even in challenging environments, has attracted significant attention. Specifically, convolutional neural networks (CNNs) have emerged as prominent tools within the realm of deep learning, enabling the development of noise-robust speech recognition systems [1]. Hence, Automatic Speech Recognition (ASR) has seen remarkable advancements due to deep learning techniques [2-6].

Furthermore, in the context of baby cry recognition, voice recognition technology offers a promising solution, leveraging advanced computational methods to automatically analyze and classify baby cries based on their acoustic features. An example of such an application is the Android-based *Madsaz Baby Cry Translator* app, which translates the cries of infants (0-3 months old) to help parents recognize various cry types and other cues. This app, available in both Indonesian and English, has been downloaded in 175 countries. User feedback suggests that the *Madsaz Baby Cry Translator* app boosts parents' confidence in childcare and enhances their responsiveness to their babies' cries. Hence, due to its potential use, it is important to improve the app by providing timely and accurate translation of baby cries, particularly in real-world, noisy environment.

Recent research has explored strategies to improve accuracy and efficiency. For instance, one study [7] demonstrated the conversion of baby cry signals into spectrograms, followed by CNN classification, achieving an impressive 99.83% accuracy. This approach effectively addresses the challenge of server workload while maintaining high performance. Another study [8] investigated the use of MFCC features extracted from baby cry signals coupled with CNNs, achieving an accuracy of 96.6%.

Unlike traditional feature extraction methods, CNNs directly extract relevant features from audio data through convolutional layers, allowing them to learn complex features in parallel during network training. This inherent learning capability makes CNNs highly adaptable and well-suited for accurate classification tasks [9]. Notably, CNNs have achieved accuracy exceeding 90% in voice detection and recognition, including applications in infant-related research [7,8,10]. Moreover, deep convolutional neural networks (DCNNs) excel at extracting informative representations from speech signals, effectively handling diverse sources of variability [11]. By strategically harnessing the strengths of CNNs, this study aims to develop a more robust and accurate baby cry recognition system that can effectively handle real-world noise conditions.

While Convolutional Neural Networks (CNNs) have established their strength in multi-label classification tasks, further advancements in feature extraction and pre-processing are crucial for optimal performance. Recurrent Neural Networks (RNNs) were prominent in this domain [11], while CNNs have achieved promising accuracy rates of 94% [12]. Another study states that the Scatter Transform-DCNN algorithm [13] demonstrates noise-robustness in classifying normal and pathological sounds. By effectively extracting features related to crying sounds through log-linear filter banks [14], CNNs have shown success in cloud-based baby cry detection (86% accuracy) [15]. While CNNs excel at capturing local spectral and temporal variations through high-level feature extraction, RNNs offer complementary strengths in capturing extended temporal contexts within audio signals [16]. This constructive collaboration led to significant accuracy improvements in polyphonic sound detection when combining CNN and RNN models [16]. Hybrid systems incorporating Restricted Boltzmann Machines (RBNs) and CNNs have also been explored for baby cry recognition, achieving 78.6% accuracy [17].

Deep Neural Networks (DNNs) have achieved significant progress in enhancing noise robustness for acoustic models, particularly regarding automatic voice recognition [10]. This task becomes challenging in noisy environments, but recent studies have shown promising results. One approach involves converting spectrograms into images, followed by dimension reduction, feature extraction, and CNN classification. This method achieved a 4.5% performance increase and a 97.4% classification success rate [18]. Another study explored a CNN architecture incorporating both short-term and long-term audio data, boosting accuracy through adaptive thresholding and early stopping [5].

Other related research explores hyperparameter optimization and network structures that can affect recognition performance while using the same input, suggesting that focusing on learning synchronization may be key in this context [23].

Convolutional Neural Networks (CNNs) come in various forms like 1D, 2D, and 3D, each offering unique strengths. For instance, Long Short-Term Memory (LSTM) networks combined with 2D CNNs have demonstrated superior performance in recognizing emotions from facial expressions, achieving 95.33% accuracy compared to 1D CNN-LSTM models [24]. Similarly, a 2D-3D CNN approach effectively captured micro-expression movements, leveraging separate networks for short-term a]. Additionally, multi-layered CNNs demonstrate a 10% noise reduction compared to traditional methods [19]. These advancements highlight the potential of DNNs for handling noise challenges in automatic voice recognition tasks.

Several approaches demonstrate success in noise-robust speech recognition, each highlighting different strengths. One method integrates MFCC and CNN, utilizing spectrograms and the Google Speech-to-Text API for noise mitigation and secure passcode generation [20]. Another study focuses on Automatic Modulation Classification (AMC) using CNNs. The bi-spectrum-based AMC method and *AlexNet* CNN enable the automatic extraction of significant features from images and subsequently assign corresponding labels, achieving a classification accuracy of 97.7% at or above 5 dB [21]. This finding aligns with research involving the utilization of CNNs to process time-frequency distributions for radio signal recognition, even at -2 dB SNR [22]. These studies suggest that DNN performance in the radio domain is not constrained by factors such as network depth or specific domains like natural language processing [23]. nd static features, improving recognition accuracy [25]. These studies highlight the effectiveness of 2D and 3D CNNs in video modelling, action recognition, and hyperspectral image analysis [26-28].

However, capturing complex textual features in human-robot interaction remains a challenge. Therefore, research on 3D CNNs for text representation continues to evolve. One recent study proposed a 3D-based approach that encodes semantic cubes, capturing local word features and sequential context. These representations are then fed into another 3D CNN to extract interactive features between sentences, resulting in final matching representations. This method achieved comparable or even better performance compared to existing state-of-the-art methods [29].

Feature extraction plays a crucial role in baby cry recognition systems, with Mel-Frequency Cepstral Coefficients (MFCC) used for their effectiveness. Research has shown that MFCC features can be successfully used to train backpropagation artificial neural networks, achieving high accuracy (98.9%) in identification [30]. Additionally, MFCCs capture feature segments sensitive to distortion, making them robust to common audio processing variations [31]. Studies comparing speaker gender recognition have highlighted the superiority of MFCCs over other methods like LPCC and PLP, achieving 99.37% accuracy with 16 coefficients [32]. Notably, MFCC outperformed LPCC in fixed-phrase speaker verification systems, demonstrating a 0% error rate [33]. The combination of MFCC feature extraction and a CNN algorithm has also shown promising results in baby cry detection, surpassing the performance of logistic regression classifiers [34]. These notions underline the value of MFCCs for accurate and robust baby cry recognition.

Several studies have explored the influence of feature extraction and noise mitigation on baby cry recognition performance. Utilizing MFCC and HMMs achieved 93.89% accuracy in noise-free environments but dropped to 58.1% with noise [35]. Conversely, a system combining MFCCs and a codebook achieved 94% accuracy in identifying different baby cry types, even with noise, by incorporating RNNs [36]. A previous study compared LSTM and GRU architectures in noise-free and noisy scenarios (5-20 dB SNR). While both models achieved high accuracy in noise-free conditions (94% with GRU), GRU performance dropped slightly to 89% with added noise [37].

Subsequently, this study proposes the use of a Convolutional Neural Network (CNN) model in the recognition system of *Madsaz Baby Cry Translator* app to address noise interference. Specifically, this research aims to compare the performance of CNNs against Recurrent Neural Networks (RNNs) to evaluate their effectiveness in handling noise and enhancing the accuracy of baby cry recognition. The paper is structured as follows: Section 1 provides an overview of the challenges associated with baby cry recognition and the potential benefits of using deep learning models to address these challenges. Section 2 reviews relevant literature on baby cry analysis, deep learning technique, and model evaluation. Section 3 details the process of baby cry data acquisition and processing. Section 4 presents the results and discussion of the comparative analysis between CNNs and RNNs. Lastly, Section 5 concludes the study, highlighting key findings and potential implications for future research.

II. Literature Review

A. *Baby Cry*

A baby's cry is more than just a sound; it is filled with emotions, movements, and expressions that serves as their primary means of communication. While often associated with negative emotions like discomfort or distress, cries can also convey hunger, fatigue, or simply a desire for interaction.

Babies tend to cry more often during the night within a 24-hour cycle [38]. Considered a form of communication, a baby's cry is classified into a speech category. In human communication, speech sometimes changes its signals to aid understanding [39-41]. Studies have broken down these sound signals into smaller units known as phonemes, utilizing diverse methods to assess each fragment within the vocal signals [40-42].

1) Dunstan baby language: Dunstan Baby Language (DBL) is a communication method tailored for understanding the cries of infants aged 0–3 months, applicable across diverse cultures and languages[1]. This language identifies five distinct variations:

- "Neh" indicates hunger, resembling the sound made when a baby tastes while breastfeeding. Recognizing "neh" involves detecting the insertion of the letter 'N' in the cry, often accompanied by actions like moving the tongue to the roof of the mouth, sucking fingers or the head, licking lips, and shaking the head from side to side.
- "Owh" signifies tiredness, akin to the sound of a yawn. Signs include restlessness, rubbing eyes, scratching, or pulling ears, and squirming while arching the body.
- "Eh" expresses the need to burp. The "eh" cry occurs when the baby's chest works hard to release gas, usually manifesting as faster and shorter in frequency as the baby attempts to burp. Other signs include a sensation of tightness in the chest, fidgety movements when laid down to rest, and ceasing to drink milk, becoming restless.
- "Eairh" denotes bloating, indicating the presence of gas in the stomach causing discomfort. This cry is prompted by stomach gas, leading to pain (colic). Other indications include the baby twitching their legs and pulling them toward the stomach, stiffness in the body, and screaming due to pain.
- "Heh" signifies discomfort. Babies might fuss because they feel uncomfortable, possibly due to a wet diaper, extreme temperatures, or other reasons. The "heh" cry tends to be breathless, sounding like an exhalation, with a notable emphasis on the letter 'H' at the beginning of the word.

2) Voice recognition of baby cry: In the field of voice recognition, two distinct domains emerge: speech recognition and speaker recognition. While speech recognition focuses on identifying the meaning encoded within spoken words, speaker recognition prioritizes identifying the individual behind the voice [43]. In the context of baby cry recognition, speech recognition algorithms strive to decode the cry itself, recognizing it as a distinct sound within the audio stream. This initial step often involves comparing the captured audio with existing databases to assess the level of sound suppression and ensure compatibility with the system's format. Once the cry is identified, the focus shifts to speaker recognition. Here, the objective is to determine the specific infant producing the cry. This crucial step relies on two key modules: feature extraction and feature matching.

Feature extraction involves collecting and quantifying specific characteristics from the cry audio by extracting a unique "fingerprint" of the sound based on various parameters like pitch, rhythm, and spectral energy distribution. This fingerprint then becomes the basis for feature matching. The extracted features are compared against a database of pre-existing cry recordings associated with individual babies [44]. This dual approach, combining speech recognition for cry identification and speaker recognition for individualization, holds significant promise for various applications.

B. Deep Learning

Artificial Intelligence (AI) embarks on a fascinating journey, simulating human intelligence within the realm of machines. From the perspective of computer science, AI revolves around "intelligent agents," devices that perceive their environment and take actions to achieve specific goals. In simpler terms, "AI" is often used when machines exhibit human-like capabilities, like learning and problem-solving. This brings machine learning under the umbrella of AI.

Machine learning, a cornerstone of modern computing, focuses on enhancing machine intelligence through extensive research. Borrowing from our natural ability to learn, this field strives to improve the accuracy of algorithms, making machines smarter and more capable. Deep learning, a subfield of machine learning, marks a significant advancement in this pursuit. Its applications have been extensively explored across diverse domains and subdomains, offering innovative solutions to complex challenges.

One key strength of deep learning lies in its ability to handle both feature extraction and classification within a single framework. This eliminates the need for manual feature engineering, which involves meticulously crafting features from raw data, often with inherent human bias. By automating this process, deep learning can handle vast numbers of layers and parameters, allowing it to learn more complex relationships within data [45]. Applying deep learning to baby cry recognition follows a similar approach. The network analyses the audio data, automatically extracts relevant features, and classifies the sounds.

1) Convolutional neural networks: Within deep learning, convolutional neural networks (CNNs) have emerged as formidable tools, captivating researchers, and practitioners alike. CNNs possess the remarkable ability to learn complex patterns directly from raw image data, eliminating the need for tedious pre-processing or feature extraction. This inherent strength makes them ideally suited for tackling diverse tasks involving two-dimensional data, such as image recognition, video analysis, and image generation.

[1] Gunawan, A. (2011). Dunstan Baby Language Indonesia. Retrieved from http://www.mommeworld.com/post/view/49/dunstan-baby-language-indonesia/.

The structure of the CNN entails several elements: firstly, an input layer for receiving and storing raw image data; secondly, a convolutional layer that enhances input features and reduces noise by utilizing kernels with weighted cells; thirdly, a pooling layer responsible for subsampling input data by dividing it into smaller regions and applying functions like maximum or average pooling to each region; and finally, a fully connected layer that connects all neurons from the previous layer to every neuron in its own layer [46]. In the CNN model, each h_{ij} hidden unit feature value is calculated as in (1) [47].

The difference between CNN and other neural network models is the convolution process within the hidden layers. The convolution process is calculated as in (1) [47]. In a convolution operation, the input is an m × M matrix. When the convolution kernel is an n × n matrix (K) and the stride is 1, the resulting matrix F has dimensions (m – n + 1) × (m – n + 1). Here, $i \in R, j \in R, k_{ij}$ denotes the value of row i and column j in convolution kernel, while x_{ij} represents the value of row i and column j in the image matrix. b_1 denotes the bias, and f is the activation function.

$$F_{ij} = f(b_1 + \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij} \times x_{ij}) \qquad (1)$$

The proposed CNN for baby cry recognition adheres to the foundational principles of CNN architecture. It comprises two convolutional layers and a single dense layer leading to the SoftMax classifier. The hidden layers utilize a Rectified Linear Unit (ReLU) activation function and employ a 50% dropout mechanism for regularization. During the initial optimization of hyper-parameters, the convolutional layers are configured with a filter size of 1x3 [23]. Fig. 1 provides a high-level overview of the CNN architecture, while Fig. 2 shows the training and testing modules in detail. Data flows through the convolutional layers, establishing connections with subsequent layers. The SoftMax function delivers probabilistic values ranging from 0 to 1, facilitating classification. The interconnected nature of CNNs simplifies both training and testing procedures by using hidden layers. Backpropagation, the fundamental algorithm in CNN, automatically computes the requisite parameters. CNN offers three primary advantages for speech recognition: location specificity, weight distribution, and pooling.

Moreover, CNN architecture incorporates these strengths to enhance noise resilience. Upper network layers can effectively handle noise due to the combination of high-level features extracted from each frequency band. Additionally, pooling reduces the number of local networks, further mitigating noise sensitivity [48].

C. Spectrogram

Spectrograms are widely employed as a common method for conducting time-frequency analysis to estimate specific signal parameters [49]. As a type of Time-Frequency Distribution, a spectrogram illustrates signal energy across both time and frequency dimensions. It is particularly useful for analyzing nonstationary signals, whose attributes fluctuate over time [49-51]. This approach efficiently captures the dynamics of such signals, which exhibit varying characteristics over time [50]. The mathematical representation of a spectrogram is outlined in (2) [51].

$$S_x(t,f) = \left|\int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f} dt\right|^2 \qquad (2)$$

The signal under analysis, denoted as x(τ), is examined within the observation window represented by w(t), t represents the time, and f the frequency.

Spectrograms serve as instrumental tools for visualizing variations within the frequency spectrum of a signal, effectively capturing dynamic changes across both temporal (e.g., audio signals, earthquake waves) and spatial dimensions (images). In the field of machine learning, spectral information derived from spectrograms frequently plays a role in revealing intricate features and patterns within the source data. Typically, the frequency spectrum of a signal is acquired through the utilization of a Fourier Transform (FT). In the case of discrete data, spectral analysis relies on the Discrete Fourier Transform (DFT), which converts a finite sequence of N complex numbers representing the signal {xn} = x0, x1..., xN−1 into a corresponding sequence of K = N complex numbers {Xk} = X0, X1..., XN−1 (3) [52]. χ_n is input sequence, X_k is the transformed input sequence, N-periodic sequence, dan k ∊ [0, N-1].

$$X_k = \sum_{n=0}^{N-1} \chi_n \, e^{-i2\pi kn/N} \qquad (3)$$

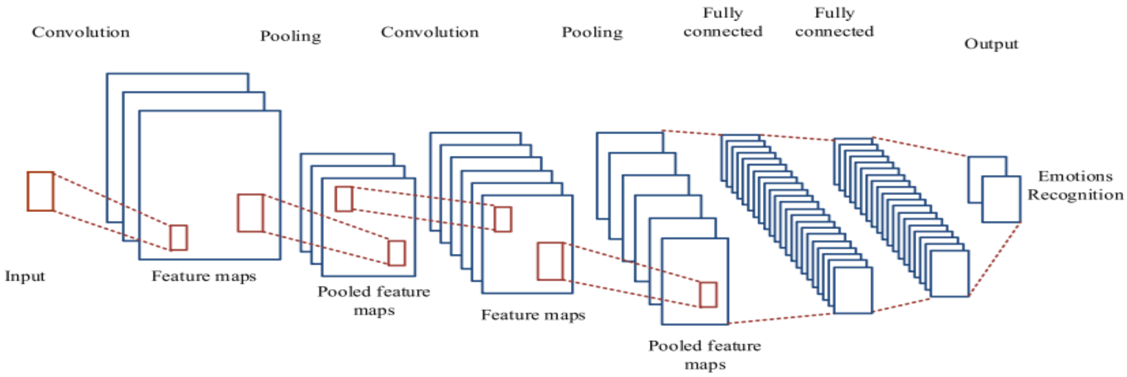


Fig. 1. CNN architecture [48].

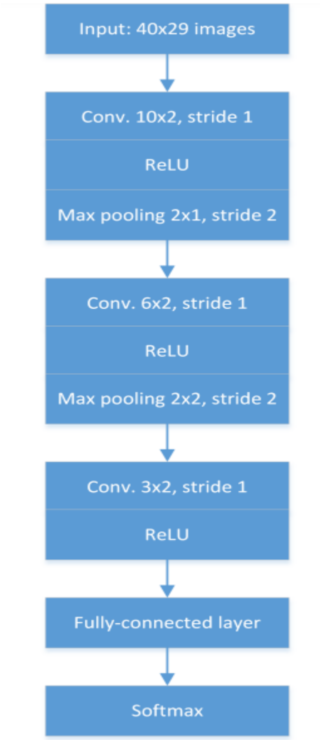


Fig. 2. CNN architecture [34].

The Short-Time Fourier Transform (STFT) generates spectrogram based on the magnitude of a discrete signal with length L. This technique leverages the Discrete Fourier Transform (DFT) to partition the signal into N segments, where N < L. This segmentation results in a complex matrix S, containing the signal's magnitude and phase across both frequency and time domains for each segment. Typically, the columns of the matrix represent the temporal dimension, while the rows correspond to different frequency bands. The chosen value of N depends on the intended spectral representation. Lower N values offer higher temporal resolution but lower frequency resolution, while higher N values yield the opposite effect. Furthermore, the STFT allows for segment selection by varying the segment index m, ranging from 0 to N-1, resulting in high temporal definition and low frequency resolution for smaller N or vice versa for higher N. Additionally, segments can be overlapped by m samples within the range of 0 and N-1, offering further control or flexibility over the desired resolution [52].

D. *Model Evaluation*

Model evaluation ensures a classification model's effectiveness. Evaluating a classification model goes beyond just checking its overall accuracy. A deeper dive into various metrics promotes understanding of its effectiveness in distinguishing between distinct categories. For both binary and multiclass classification problems, the confusion matrix holds a central position as an indispensable evaluation tool [53]. Table I displays the structure of the tool in the field of binary classification as the essence of the confusion matrix, providing an illustration of the model's performance [54].

TABLE I. CONFUSION MATRIX SCHEME

	Predicted Class	
True Class	True Positive (TP)	True Negative (TN)
	False Positive (FP)	False Negative (FN)

Classification models generate four key values: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). Each value provides insights into the model's ability to distinguish between various categories. TP denotes the number of instances correctly identified and predicted as positive, while FP indicates the number of instances incorrectly identified as positive when they are actually negative. FN signifies the count of instances incorrectly identified as negative when they are positive, while TN signifies the count of instances accurately identified and predicted as negative. Performance metrics commonly employed in classification tasks including the accuracy value (ACC) (4), precision (P), representing the probability of a case being predicted as positive when it truly belongs to the positive category (5), F-score value (6), and recall (7).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$P = \frac{TP}{TP+FP} \tag{5}$$

$$F_{score} = 2X\frac{PxSn}{P+Sn} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

III. BABY CRY DATA ACQUISITION AND PROCESSING

This research explores the performance of a Convolutional Neural Network (CNN) for baby cry recognition, comparing it to an existing Recurrent Neural Network (RNN) approach previously implemented in Madsaz Baby Cry Recognition dataset [37]. The dataset comprises 175 records data categorized into five distinct cry types, with balanced representation in both training (80%) and validation (20%), respectively. To simulate real-world noise interference, the study integrated the original baby cry signals with Gaussian noise. This type of noise was chosen because it closely resembles background noise commonly encountered in real environments. To control the intensity of the noise, the signal-to-noise ratio (SNR) was varied between 5 and 20 dB, representing a range from moderate to significant noise levels. Examples of the noise-free and noise-added cry signals are presented in Table II for visual comparison.

TABLE II. BABY CRY DATA

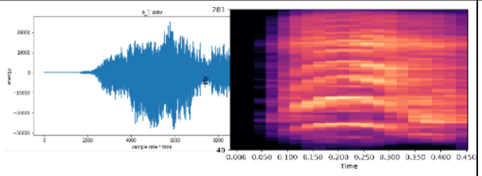
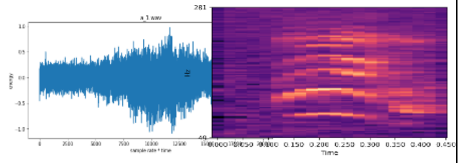
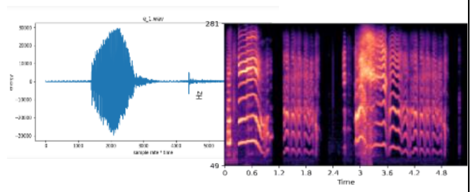
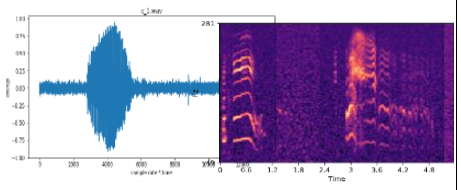
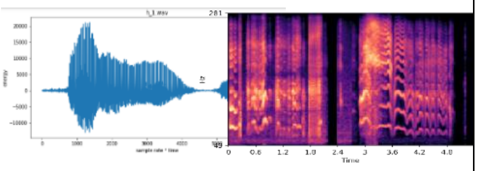
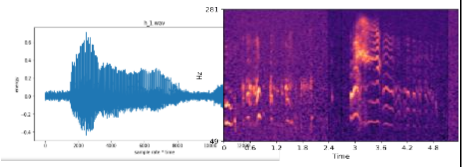
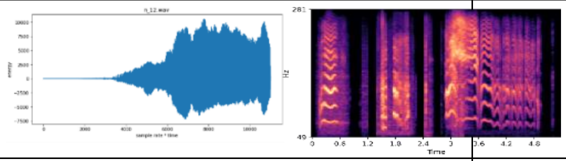
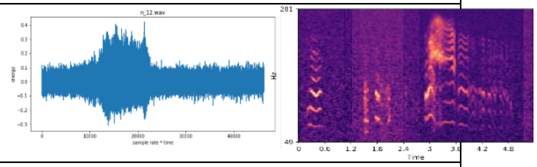
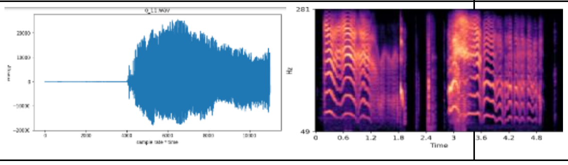
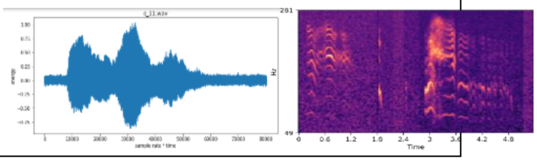
Classification	*Signal and spectrogram without noise-added data*		*Signal and spectrogram with noise-added data*
Eairh			
Eh			
Heh			
Neh			
Owh			

Fig. 3 illustrates the proposed CNN architecture, featuring three stacked convolutional layers followed by a max-pooling layer and a dropout layer (set to 0.5). Each convolutional layer utilizes various filter sizes (64, 128, and 256) with a 3x3 kernel, employing the same padding and ReLU activation function for nonlinearity. The extracted features are then flattened and fed into a fully connected layer with 512 units, again using ReLU activation. Another dropout layer (0.5) precedes the final output layer with a SoftMax activation function, capable of assigning probabilities to each of the five cry categories.

The model is trained with an input size of 64x64, utilizing the Adam optimizer and the sparse categorical cross-entropy loss function. During training, batches of 64 samples are processed for 100 epochs. The performance is evaluated using various metrics, including precision, recall, F1 score, and overall accuracy.

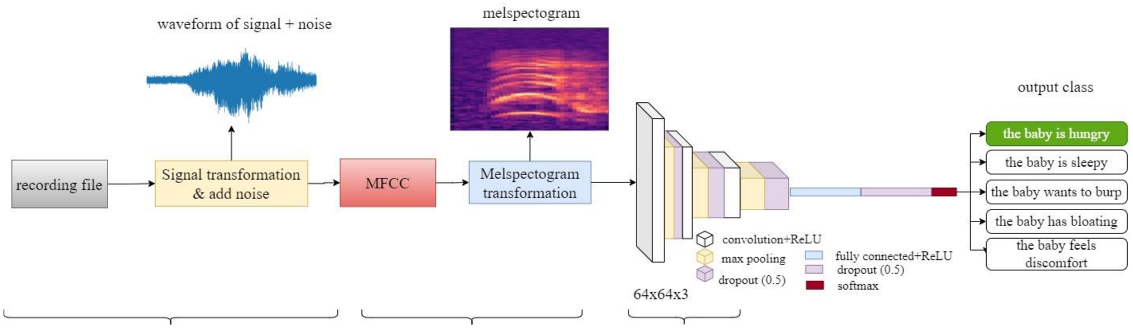


Fig. 3. CNN architecture.

IV. Results and Discussions

The results highlight the remarkable robustness of the CNN method to noise interference, primarily due to its inherent convolutional architecture. This feature allows the CNN to extract more comprehensive features from the data, particularly when sound signals are converted into visual representations. This translates to more robust and generalizable results, even in the presence of noise. The CNN achieved 94% accuracy with noise-free data and maintained a 91% accuracy when noise was introduced. Fig. 4 and 5 compare the training and validation accuracy under both noise-free and noise-added conditions, visually demonstrating the CNN's resilience. Additionally, Fig. 6 provides a confusion matrix for the CNN when applied to noise-added data, offering further insights into its performance.

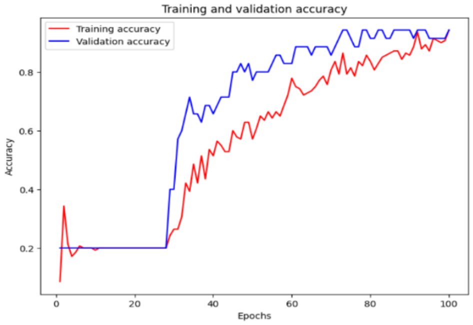


Fig. 4. Graph of training and validation accuracy without noise.

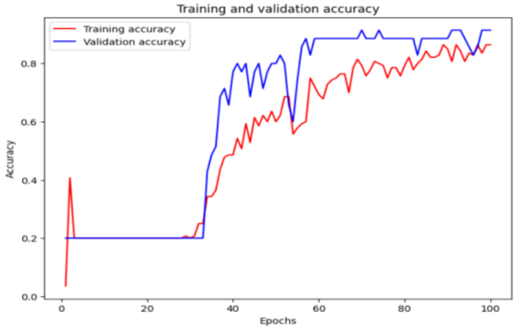
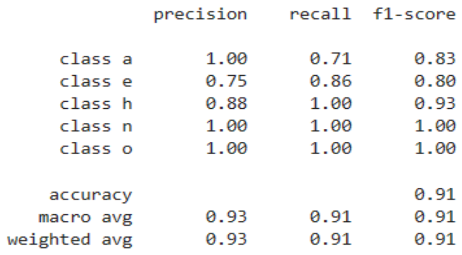


Fig. 5. Graph of training and validation accuracy with noise.



```
              precision    recall  f1-score

     class a       1.00      0.71      0.83
     class e       0.75      0.86      0.80
     class h       0.88      1.00      0.93
     class n       1.00      1.00      1.00
     class o       1.00      1.00      1.00

    accuracy                           0.91
   macro avg       0.93      0.91      0.91
weighted avg       0.93      0.91      0.91
```

Fig. 6. Evaluation mode using methods with noise-added data.

Building upon the above result of superior CNN performance, further analysis delves into the specific advantages it holds compared to Recurrent Neural Networks (RNNs) like Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) for baby cry recognition using Mel spectrograms. While all three models process the Mel spectrogram data, their fundamental approaches differ significantly, impacting their resilience to noise and recognition accuracy. The goal is to assess their robustness against noise in the input data. A detailed comparison of the performance between RNN and CNN methods can be referred to Table III. Precision considers only positive predictions, which can result in high precision even if there are numerous negative instances misclassified as false negatives. Recall measures the model's ability to identify all positive instances within the dataset. Conversely, accuracy considers all predictions, both positive and negative, thus providing a more comprehensive reflection of overall model performance. This is particularly useful when the dataset in this research is balanced, i.e. the number of positive and negative instances is almost equal [55].

TABLE III. Comparative Performace of RNN and CNN Methods in Baby Cry Recognition System

	Precision	*Recall*	*F1 Score*	*Accuracy*
RNN-GRU (noise-free data)	91%	89%	88%	89%
RNN-GRU (noise-added data)	91%	89%	88%	89%
RNN-LSTM (noise-free data)	91%	89%	88%	89%
RNN-LSTM (noise-added data)	96%	77%	77%	77%
CNN (noise-free data)	96%	94%	94%	94%
CNN (noise-added data)	93%	91%	91%	91%

Table III demonstrates the findings revealing that CNNs consistently outperformed GRUs and LSTMs with higher accuracy in both noise-free and noise-added settings. This advantage stems from CNNs' ability to extract spatial features from the Mel spectrograms. These features capture the patterns and characteristics of baby cries, allowing the CNN to recognize them more accurately and sustain its performance level.

While GRU and LSTM are powerful for sequential data, they face challenges when applied to baby cry recognition using Mel spectrograms, which are spatial representation. This study shows that RNNs rely on connections throughout the data sequence, leading to the vanishing gradient problem where information gets lost over time. Moreover, RNNs need to be adapted for spatial data, despite being designed for sequences. A detailed comparison between the two RNN models revealed that GRU outperformed LSTM, especially in noise-added settings. The result of analysis indicates that the benefit of GRU lies in its simpler and more efficient architecture with only two gates, the reset and update gates. This streamlined design helps mitigate the vanishing gradient problem and balances model complexity with the ability to understand the context of the Mel spectrogram.

The reset gate in GRU plays a role in preventing vanishing gradients by allowing the model to selectively "forget" less relevant information, including noise, preventing it from accumulating and impacting the recognition process. In contrast, LSTM's three gates, including the input gate, forget gate, and output gate, retain information for longer durations due to its extended memory. While this can be beneficial for some tasks, it also increases the risk of preserving noise, hindering accurate recognition. However, it is important to remember that the selection between GRU and LSTM also hinges on the specific attributes of the dataset and the requirements of the task. For recognizing baby cries from Mel spectrograms, GRU's simpler architecture and ability to handle noise seem to offer an advantage.

GRU consistently maintains high precision, recall, F1-Score, and accuracy rates (91%, 89%, 88%, and 89%) across data in both noise-free and noise-added settings. This highlights the model's robust and consistent ability to accurately identify baby cries, minimizing both false positives and negatives. In comparison, LSTM delivered similar results on noise-free data. However, its performance noticeably decreased by approximately 8.79% for precision, 13.48% for recall, 12.50% for F1-score, and 13.48% for accuracy with the addition of noise. This underscores the LSTM's higher susceptibility to noise interference in classifying baby crying sounds. On the other hand, the CNN model exhibits excellent performance in both scenarios, outperforming the GRU and LSTM. The performance of the CNN model decreased by around 3.13% for precision, 3.19% for recall, 3.19% for F1-score, and 3.19% for accuracy in the presence of noise. Despite the decrease, the CNN model remained superior in identifying baby crying sound patterns compared to both GRU and LSTM in noise-free data settings. The results of this study also strengthen [56], which shows that CNN has better performance than other deep learning models in classifying baby crying sounds using spectrogram features.

Research combining CNN and RNN models [57] provides an accuracy of around 94%, but the model does not accommodate recognition in the presence of noise. Overall, the CNN exhibits superiority in robustness in this study, making it a valuable tool for baby cry recognition in real-world settings with potential noise interference. The performance of CNNs, particularly in noisy environments, has significant implications for practical applications like baby monitoring systems. The ability to accurately recognize cries despite background noise can enhance safety and responsiveness, contributing to improved care and well-being.

V. Conclusion

This study explored the potential of deep learning approaches for enhanced baby cry recognition while mitigating noise interference. Using comparative analysis, two architectures were evaluated, *i.e.*, Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). The evaluation was conducted with both noise-free and noise-added data (SNR 5-20 dB), revealing the superior robustness of CNNs against noise.

The advantage comes from CNNs' ability to extract noise-resistant features from the Mel spectrogram representation of audio signals. These features, such as spectral energies and formant frequencies, are crucial for cry recognition and remain relatively intact even in noisy environments. In contrast, RNNs, particularly LSTMs, might capture irrelevant noise information due to their longer memory retention, leading to performance degradation.

The findings demonstrate that the CNN achieved an impressive 94% accuracy on noise-free data, maintaining an outstanding 91% accuracy on noise-added data. This minimal performance drop displays the significant advantage of CNNs in real-world scenarios with potential noise interference. Further analysis revealed that CNNs excel in understanding the spatial structure of data, crucial for analyzing Mel spectrograms. Their inherent flexibility in handling image-like representations, regardless of noise, contributes to a stable and accurate recognition process. In conclusion, this study highlights the potential of CNNs for robust baby cry recognition, particularly in noisy environments. The ability to extract noise-resistant features and utilize spatial information positions CNNs as a valuable tool for applications requiring accurate cry detection in real-world settings.

References

[1] Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., Jin, W., & Schuller, B. (2018). Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. ACM Transactions on Intelligent Systems and Technology, 9(5), 2–28. doi: 10.1063/5.0032382.

[2] Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Trans. Audio. Speech. Lang. Processing, 20, 30–42. doi: 10.1109/TASL.2011.2134090.

[3] Amodei, D., et al. (2016). Deep speech 2: End-to-end Speech Recognition in English and Mandarin. In Proceedings of the International Conference on Machine Learning (ICML'16), 173–182.

[4] Saon, G., Sercu, T., Rennie, S., & Kuo, H.-K. J. (2016). The IBM 2016 English conversational telephone speech recognition system. In Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'16), 7–11.

[5] Jeong, I., Lee, S., Han, Y., & Lee, K. (2017). Audio Event Detection using Multiple-Input Convolutional Neural Network. Dcase 2017, 1(November), 2–5.

[6] Nagajyothi, D., & Siddaiah, P. (2018). Speech recognition using convolutional neural networks. Int. J. Eng. Technol., 7(4.6 Special Issue 6), 133–137. doi: 10.14419/ijet.v7i4.6.20449.

[7] Chang, C. Y., & Tsai, L. Y. (2019). A CNN-Based Method for Infant Cry Detection and Recognition (Vol. 927). Springer International Publishing.

[8] Zabidi, A., et al. (2017). Detection of Asphyxia in Infants using Deep Learning Convolutional Neural Network (CNN) Trained on Mel Frequency Cestrum Coefficient (MFCC) Features Extracted from Cry Sounds. Journal of Fundamental and Applied Science, 9(3S), 768–778.

[9] Bashar, D. A. (2019). Survey on Evolving Deep Learning Neural Network Architectures. Journal of Artificial Intelligence and Capsule. Networks, 2019(2), 73–82. doi: 10.36548/jaicn.2019.2.003.

[10] Seltzer, M. L., Yu, D., & Wang, Y. (2013). An Investigation of Deep Neural Networks for Noise Robust Speech Recognition. ICASSP, IEEE International Conference on Acoustic, Speech, and Signal Process. - Proc., 7398–7402. doi: 10.1109/ICASSP.2013.6639100.

[11] Song, G., Wang, Z., Han, F., Ding, S., & Iqbal, M. A. (2018). Music auto-tagging using deep Recurrent Neural Networks. Neurocomputing, 292, 104–110. doi: 10.1016/J.NEUCOM.2018.02.076.

[12] Song, G., Wang, Z., Han, F., Ding, S., & Gu, X. (2020). Music auto-tagging using scattering transform and convolutional neural network with self-attention. Applied Soft Computing Journal, 96, 106702. doi: 10.1016/J.ASOC.2020.106702.

[13] Souli, S., Amami, R., & Ben Yahia, S. (2021). A Robust Pathological Voices Recognition System Based on DCNN and Scattering Transform. Appl. Acoustic., 177, 107854. doi: 10.1016/j.apacoust.2020.107854.

[14] Xie, J., Long, X., Otte, R. A., & Shan, C. (2021). Convolutional Neural Networks for Audio-Based Continuous Infant Cry Monitoring at Home. IEEE Sens. J., 21(24), 27710–27717. doi: 10.1109/JSEN.2021.3123906.

[15] Zhang, X., Zou, Y., & Liu, Y. (2018). AICDS: An Infant Crying Detection System Based on Lightweight Convolutional Neural Network (Vol. 10970 LNCS). Springer International Publishing.

[16] Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. IEEE/ACM Trans. Audio, Speech, Lang. Process., 25(6). doi: 10.1109/TASLP.2017.2690575.

[17] Yong, B. F., Ting, H. N., & Ng, K. H. (2019). Baby Cry Recognition using Deep Neural Networks. IFMBE Proc., 68(3), 809–813. doi: 10.1007/978-981-10-9023-3_147.

[18] Ozer, I., Ozer, Z., & Findik, O. (2017). Noise Robust Sound Event Classification with Convolutional Neural Network. Neurocomputing. doi: 10.1016/j.neucom.2017.07.021.

[19] Qian, Y., Bi, M., Tan, T., & Yu, K. (2016). Very Deep Convolutional Neural Networks for Noise. IEEE/ACM Trans. Audio, Speech, Lang. Process., 24(12), 2263–2276.

[20] Chandankhede, P. H., Titarmare, A. S., & Chauhvan, S. (2021). Voice recognition-based security system using convolutional neural network. Proc. - IEEE 2021 International Conference on Computing, Communication, and Intelligent System (ICCCIS) 2021, pp. 738–743. doi: 10.1109/ICCCIS51004.2021.9397151.

[21] Li, Y., Shao, G., & Wang, B. (2019). Automatic Modulation Classification Based on Bispectrum and CNN. Proc. 2019 IEEE 8th Jt. Int. Inf. Technol. Artif. Intell. Conf. ITAIC 2019, 311–316. doi: 10.1109/ITAIC.2019.8785692.

[22] Zhang, M., Diao, M., & Guo, L. (2017). Convolutional Neural Networks for Automatic Cognitive Radio Waveform Recognition. IEEE Access, 5, 11074–11082. doi: 10.1109/ACCESS.2017.2716191.

[23] West, N. E., & O'Shea, T. (2017). Deep Architectures for Modulation Recognition. doi: 10.1109/DySPAN.2017.7920754.

[24] Zhao, J., Mao, X., & Chen, L. (2019). Speech Emotion Recognition using Deep 1D & 2D CNN LSTM Networks. Biomed. Signal Process. Control, 47, 312–323. doi: 10.1016/j.bspc.2018.08.035.

[25] Wang, L., Jia, J., & Mao, N. (2020). Micro-Expression Recognition Based on 2D-3D CNN, 3152–3157.

[26] Alayrac, J., Carreira, J., & Zisserman, A. (2019). The Visual Centrifuge: Model-free Layered Video Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2457–2466.

[27] Jiang, S., Qi, Y., Zhang, H., Bai, Z., Lu, X., & Wang, P. (2020). D3D: Dual 3D Convolutional Network for Real-time Action Recognition. IEEE Trans. Ind. Inf, 17(7), 4584–4593.

[28] Liu, X., Sun, Q., Meng, Y., Fu, M., & Bourennane, S. (2018). Hyperspectral Image Classification Based on Parameter-optimized 3D-CNNs Combined with Transfer Learning and Virtual Samples. Remote Sens., 10(9). doi: 10.3390/rs10091425.

[29] Lu, W., Yu, R., Wang, S., Wang, C., Jian, P., & Huang, H. (2021). Sentence Semantic Matching Based on 3D CNN for Human Robot Language Interaction. ACM Trans. Internet Technol., 21(4). doi: 10.1145/3450520.

[30] On, C. K., Pandiyan, P. M., Yaacob, S., & Saudi, A. (2006). Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition. 2006 Int. Conf. Comput. Informatics, ICOCI '06, 2, 2–6. doi: 10.1109/ICOCI.2006.5276486.

[31] Yuan, X. C., Pun, C. M., & Chen, C. L. P. (2015). Robust Mel-Frequency Cepstral Coefficients Feature Detection and Dual-tree Complex Wavelet Transform for Digital Audio Watermarking. Inf. Sci. (Ny)., 298, 159–179. doi: 10.1016/j.ins.2014.11.040.

[32] Yücesoy, E., & Nabiyev, V. V. (2014). Comparison of MFCC, LPCC and PLP Features for The Determination of a Speaker's Gender, 321–324.

[33] Yang, H., Deng, Y., & Zhao, H. (2019). A Comparison of MFCC and LPCC with Deep Learning for Speaker Recognition. ACM, 160–164.

[34] Lavner, Y., Cohen, R., Ruinskiy, D., & Ijzerman, H. (2016). Baby Cry Detection in Domestic Environment using Deep Learning. ICSEE Int. Conf. Sci. Electr. Eng. doi: 10.1109/ICSEE.2016.7806117.

[35] Sidiq, M., B. W, T. A., & Sa'adah, S. (2015). Desain dan Implementasi Voice Command Menggunakan Metode MFCC dan HMMs. e-Proceeding of Engineering, 2(1), 1362–1373.

[36] Renanti, M. D., Buono, A., & Kusuma, W. A. (2013). Infant Cries Identification by using Codebook as Feature Matching, and MFCC as Feature Extraction. J. Theor. Appl. Inf. Technol., 56(3), 437–442.

[37] Renanti, M. D., Buono, A., Priandana, K., & Wijaya, S. H. (2023). Noise-Robust in the Baby Cry Translator Using Recurrent Neural Network Modelling. J. Theor. Appl. Inf. Technol., 101(2), 815–826.

[38] Barr, R. G., Kramer, M. S., Boisjoly, C., McVey-White, L., & Plesst, I. B. (1988). Parental Diary of Infant Cry and Fuss Behaviour. Arch. Dis. Child., 63, 380–387.

[39] Rahim, M. G. (1994). Artificial Neural Network for Speech Analysis/Synthesis. London: Chapman&Hall.

[40] Ackenhusen, J. G. (2001). Real-time Signal Processing: Design and Implementation of Signal Processing Systems. New Jersey: Prentice-Hall, Upper Saddle River.

[41] Quatueri, T. E. (2002). Discrete-time Speech Signal Processing: Principles and Practice. Prentice Hall Signal Processing Series.

[42] Gold, B., & Morgan, N. (2000). Speech and Audio Signal Processing: Processing and Perception of Speech and Music. New York: John Wiley & Sons, Inc.

[43] Kurniawan, W. (2016). Identifikasi Speech Recognition Manusia dengan Menggunakan Average Energy dan Silent Ratio Sebagai Feature Extraction Suara pada Komputer. Biospecies, 9(1), 1–6.

[44] Gupta, D., C, M. R., Manjunath, N., & PB, M. (2012). Isolated Word Speech Recognition Using Vector Quantization (VQ). Int. J. Adv. Res. Comput. Sci. Softw. Eng., 2(5).

[45] Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. In Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA (pp. 1–6). doi: 10.1109/ICCUBEA.2018.8697857.

[46] Hao, X., Zhang, G., & Ma, S. (2016). Deep Learning. International Journal of Semantic Computing., 10(3), 417–439. doi: 10.1142/S1793351X16500045.

[47] Z. Zhang, X. Cui, Q. Zheng, and J. Cao, "Land use classification of remote sensing images based on convolution neural network," Arab. J. Geosci., vol. 14, no. 4, 2021, doi: 10.1007/s12517-021-06587-5.

[48] Pawar, M. D., & Kokate, R. D. (2021). Convolution Neural Network based Automatic Speech Emotion Recognition using Mel-Frequency Cepstrum Coefficients. Multimed. Tools Appl., 80, 15563–15587. doi: 10.1007/s11042-020-10329-2.

[49] Cohen, L. (1995). Time-Frequency Analysis. Upper Saddle River, NJ: Prentice-Hall.

[50] Kasim, R., Abdullah, A. R., Selamat, N. A., Abidullah, N. A., & Zawawi, T. N. S. T. (2015). Lead Acid Battery Analysis Using Spectrogram. Appl. Mech. Mater., 785, 692–696. doi: 10.4028/www.scientific.net/amm.785.692.

[51] Norddin, N., et al. (2013). High Voltage Insulation Surface Condition Analysis using Time Frequency Distribution, 7(7), 833–841.

[52] Garcia, M. A., & Destefanis, E. A. (2018). Spectrogram Prediction with Neural Networks, (1), 42–51.

[53] Kulkarni, A., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy.

[54] Demir, F. (2022). Deep autoencoder-based automated brain tumor detection from MRI data. Artif. Intell. Brain-Computer Interface, 317–351. doi: 10.1016/B978-0-323-91197-9.00013-8.

[55] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020.

[56] Y. C. Liang, I. Wijaya, M. T. Yang, J. R. Cuevas Juarez, and H. T. Chang, "Deep Learning for Infant Cry Recognition," Int. J. Environ. Res. Public Health, vol. 19, no. 10, 2022, doi: 10.3390/ijerph19106311.

[57] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN - RNN," J. Phys. Conf. Ser., vol. 1528, no. 1, 2020, doi: 10.1088/1742-6596/1528/1/012019.