IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 August 2025, accepted 8 September 2025, date of publication 12 September 2025,
date of current version 19 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3609459

RESEARCH ARTICLE

Unveiling Hidden Patterns in Infant Cry Audio: A Multi-Feature Vision Transformer Approach With Explainable AI

AHMAD HASASNEH[1], SARI MASRI[1], AND CHAKIB TADJ[2]
[1]Department of Natural, Engineering and Technology Sciences, Faculty of Graduate Studies, Arab American University, Ramallah, Palestine
[2]Department of Electrical Engineering, École de Technologie Supérieur, Université du Québec, Montreal, QC H3C 1K3, Canada

Corresponding author: Ahmad Hasasneh (ahmad.hasasneh@aaup.edu)

ABSTRACT The early detection and diagnosis of neonatal problems are critical to ensuring that an infant receives timely medical attention, which greatly enhances health outcomes. In this study, we propose a novel deep learning framework that listens to an infant's cry to identify and diagnose six separate conditions: one being healthy and the other five comprising sepsis, respiratory distress syndrome, jaundice, hyperbilirubinemia, and vomiting. The study utilizes a rich dataset of infant cry recordings from which key acoustic features such as spectrograms, Mel-spectrograms, and Gammatone Frequency Cepstral Coefficients (GFCCs) are extracted. A sophisticated Vision Transformer (ViT) model was developed and meticulously fine-tuned to achieve an impressive 99% classification accuracy through cross-validation. To enhance the model's interpretability, powerful explainable artificial intelligence (XAI) methods such as LRP, LIME, and attention imaging were implemented to clarify the reasoning behind the model's outputs. Through cross-validation tests, the model's trustworthiness and extensive generalizability were assessed. The findings underscore the promising capabilities of employing transformer-based deep learning frameworks along with multimodal acoustic features and explanatory methods to improve cry analysis in infants and their usable scopes in pediatric medicine.

INDEX TERMS Infant cry classification, vision transformer (ViT), explainable AI (XAI), layer-wise relevance propagation (LRP), local interpretable model-agnostic explanations (LIME), feature extraction, spectrogram, Gammatone frequency cepstral coefficients (GFCC), mel-Spectrogram, multi-feature audio representation.

I. INTRODUCTION

Neonatal mortality remains one of the most challenging public health problems to tackle. The under-five mortality rate has decreased to approximately 38 deaths per 1,000 live births, whereas deaths of neonates within the first 28 days of life are estimated at 17 per 1,000. An estimated 47% of the total deaths of children under five in 2019 were neonatal deaths, and this approximate triad can be observed within the first week of life [1], [2]. Sepsis and Respiratory Distress Syndrome (RDS) are the two most predominant causes of neonatal death. Approximately 20% to 36% of deaths are attributable to sepsis [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Cheolsoo Park.

In addition, the mortality rate of RDS can exceed 30% without critical interventions [4]. Furthermore, neonatal jaundice is a common condition, affecting about 60% of term and up to 80% of preterm neonates worldwide, and can lead to severe neurological complications if left untreated [5]. Hyperbilirubinemia, often associated with jaundice, affects approximately 1.1 million neonates annually and can lead to severe neurologic injury or death if left untreated [5], [6]. Vomiting in neonates may be indicative of severe underlying disease, further contributing to neonatal morbidity and mortality [7]. Current diagnostic techniques for these conditions rely on blood tests, X-rays, and cultures, which are often non-diagnostic and resource-intensive in low-income settings [8].

The potential for diagnosing health conditions from the sound of an infant's cry has been studied for quite some time.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

According to an early study, crying neonates with various pathologies had different cry acoustics compared to healthy infants [9]. Currently, advances in machine learning (ML) and deep learning (DL) have improved the processing of infant cry signals, which holds great promise for non-invasive early detection and diagnosis of neonatal diseases. However, even with these advances, most studies focus on differentiating healthy cries from a limited number of pathologies, usually involving fewer pathologies and rarely incorporating multiple acoustic features into a single model. In addition, there are still significant gaps in the performance and explainability of deep learning models that can be trusted in a clinical setting, which may hinder adoption [10], [11].

This research aims to bridge the existing gaps by developing an advanced deep learning-based neonatal cry diagnostic system (NCDS) designed to accurately distinguish between six conditions: Healthy, Sepsis, RDS, Jaundice, Hyperbilirubinemia, and Vomiting. The contributions of this research include the following:

1) An innovative multimodal framework that combines spectrograms, Mel spectrograms, and Gammatone Frequency Cepstral Coefficients (GFCC) derived from infant cries to enhance diagnostic accuracy.

2) The application of cutting-edge Vision Transformer (ViT) models, trained through rigorous cross-validation to ensure the reliability and robustness of the model.

3) The application of Explainable Artificial Intelligence (XAI) methods like Layer-wise Relevance Propagation (LRP), Local Interpretable Model-Agnostic Explanations (LIME), and attention maps adds reasoning capabilities regarding how the model made its predictions.

4) A novel combination of LRP and LIME with a multi-feature ViT model applied for the first time in infant-cry diagnostics, delivering pixel-level explanations for each individuals prediction.

The study results emphasize the promise of enhanced non-invasive diagnostic techniques for neonates by synergizing multi-modal acoustic features with interpretable deep learning transformers, which may help decrease the global neonatal mortality rate.

II. LITERATURE REVIEW

The analysis of infant cries has become a vital non-invasive tool for early medical diagnostics, capturing the attention of researchers for decades. Similar to many fields, it has evolved from simple listening and observation to more sophisticated processing systems that leverage machine learning (ML) and deep learning (DL) algorithms.

Previous studies have demonstrated that specific acoustic features of infant cries could indicate various health condition [9]. Building on this idea, later research works have mainly employed classical ML techniques. Early and commonly used acoustic features include Mel-frequency cepstral coefficients (MFCC), Linear-frequency cepstral coefficients (LFCC), and Bark frequency cepstral coefficients (BFCC) [12], [13]. For instance, several studies [14], [15] utilized MFCC features combined with traditional ML classifiers such as support vector machine (SVM) and K-nearest neighbor rule (KNN) along with Gaussian mixture models (GMMs), achieving varying levels of success with classification accuracies ranging from 71% to 78% for pathological infant cries.

With the rise of deep learning methods, convolutional neural networks (CNNs) started to outperform classical ML approaches, particularly in cry classification tasks [16], [17], [18], [19]. CNNs showed remarkable performance with Spectrogram, Mel-spectrograms, and GFCCs features, treating these as image-based audio data [16], [17]. Also, the authors in [18] used spectrogram images with CNNs for classification tasks, achieving an accuracy of nearly 89% in distinguishing cries for pain, hunger, and sleepiness, demonstrating a significant improvement in the overall classification accuracy. Similarly, the researchers in [19] achieved even higher classification accuracies, surpassing 97%, by using deep learning models with GFCCs and spectrograms.

Other recent studies have taken a step further by proposing the use of transformer models, particularly ViTs, to improve the classification accuracy due to their ability to capture complex patterns and dependencies in spectrogram data [20]. More precisely, the researchers in [20] have highlighted that ViTs outperformed traditional CNNs in large-scale image recognition tasks, which led to the adaptation of these ideas for audio classification. In addition, the authors in [21] have boldly applied transformers to audio classification, reporting significant performance improvements compared to CNNs-based approaches. Furthermore, in [22], the Audio Spectrogram Transformer (AST), a variant of ViT designed to process raw audio data, was recently applied to classify three different classes and achieved an impressive 97% accuracy.

On the other hand, the application of XAI methods to transformer architectures has significantly improved the trustworthiness and explainability of audio diagnostics. Techniques such as LRP, LIME, and attention mechanisms have been used to provide explanations for the model's decisions [23], [24]. Recently, the authors in [25] have built upon these advanced interpretability techniques, achieving significant performance improvements (96.33% accuracy) and enhanced interpretability in infant cry diagnostics using ViT models on GFCCs and spectrogram representations.

Despite these advancements, there is a clear gap in the literature. One can see that most current research works have focused on a limited binary scope or multiclass settings, such as healthy versus pathological cries, or comparing minor conditions like RDS and sepsis [11]. Few studies have explored multiclass classification involving multiple common neonatal conditions. Moreover, while automated machine learning (AutoML) approaches for combining multimodal features (such as Spectrogram, Mel-spectrograms, and GFCCs) showed promising results, however, very few studies have attempted to incorporate these features into a single cohesive model.

This research seeks to address existing gaps by building upon prior work, including [25] and developing a more sophisticated multiclass diagnostic framework that

automatically diagnoses six neonatal conditions: Healthy, Sepsis, RDS, Jaundice, Hyperbilirubinemia, and Vomiting. The research design incorporates advanced multimodal ViT architectures combined with acoustic features to enhance both accuracy and interpretability. Furthermore, the study employs robust explanatory techniques such as LRP, LIME, and attention mechanisms to enhance understanding of decision-making process, fostering trust and facilitating the use of the model in clinical settings.

As highlighted throughout this review, significant progress has been made in cry-based infant diagnostic systems, from traditional feature extraction methods to deep learning and transformer-based models. However, more attention is needed on multimodal feature fusion and improving model explainability. This is especially critical in under-resourced regions, where accurate and non-invasive neonatal diagnostic tools could have a profound impact on reducing infant mortality and improving early intervention outcomes.

III. MATERIAL AND METHODS

The proposed methodology for the classification of the infant cry audio signals is systematic, as shown in Figure 1. First, the audio signals were processed into three images representing audio features: the spectrogram, associated with the Mel-spectrogram, and GFCCs. These features were normalized and combined into a feature image, which was represented in an RGB color channel. This image was used to identify six pathologies, as shown in Figure 2. In the next step, the images were divided into training and test sets, which were fed into the ViT model for evaluation. Cross-validation was utilized to improve generalization. Finally, model decisions were interpreted using XAI approaches, such as LIME and LRP, to maintain model transparency and clinical relevance.

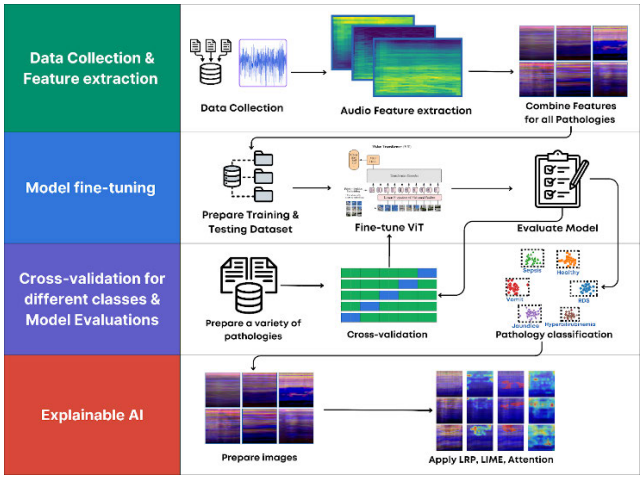

FIGURE 1. The workflow of the proposed methodology for classifying infant cries.

A. DATASET DESCRIPTION

The audio data in this research was collected from Al-Raee and Al-Sahel hospitals in Lebanon, as well as from Saint-Justine Children's Hospital in Montreal, Canada. This dataset has already been utilized in several previous studies [10], [11], [15], [19], [22], [25] and achieved promising results. It contains recordings of crying newborns from diverse ethnic backgrounds, ranging in age from 1 to 53 days, as summarized in Table 1. Each infant was recorded five times, with each recording lasting approximately 90 seconds. Data collection was conducted using a two-channel Olympus digital recorder with 16-bit resolution and a 44,100 Hz sampling rate, positioned 10-30 cm from the infant. This dataset comprises 17 infants, of whom approximately 65% (11 infants) are male and 35% (6 infants) are female. Six pathological conditions are represented: Healthy, Sepsis, Respiratory Distress Syndrome (RDS), Jaundice, Hyperbilirubinemia, and Vomiting. The audio files were segmented and labeled using the WaveSurfer program (version 1.8.8). To ensure fair and unbiased model training, the dataset was randomly down-sampled to match the size of the smallest class (929 records), so that each experimental scenario contained an equal number of samples per class, as shown in Figure 2.

Although the corpus contains recordings from only 17 neonates (approximately 1.3 hours of crying), it remains a widely accepted benchmark. To reduce the risk of overfitting, we implemented subject-independent five-fold cross-validation, balanced the classes through down-sampling to 929 records, and applied extensive data augmentation techniques.

TABLE 1. Demographic details of the dataset.

Demographic Factors	Details
Race	Asian, Arabic, African, Asian, White, Latino, Native Hawaiian, and Quebec
Origin	Canada, Algeria, Palestine, Bangladesh, Haiti, Portugal, Syria, Lebanon, and Turkey
Gender	6 females and 11 males
Weight	0.98kg to 5.2 kg
Ages	1 day to 53 days

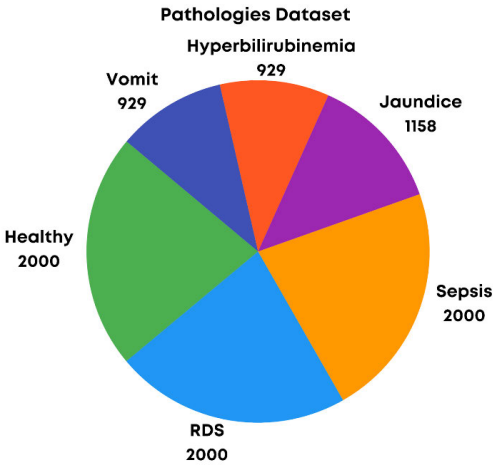

FIGURE 2. Class distribution of the dataset across the six neonatal conditions.

B. FEATURE EXTRACTION

In context of audio signal analysis of infant crying, the first step of transforming the infant cry signals into their audio features is called feature extraction, which in this case includes the extraction of a spectrogram. As described in Figure 3, three audio features were extracted: the spectrogram, the Mel-spectrogram, and GFCCs. The spectrogram is derived from the Short-Time Fourier Transform (STFT) and provides a visual representation of how the frequency spectrum of a signal changes over time [26]. GFCC features are extracted using a Gammatone filterbank (GF) followed by a Discrete Cosine Transform (DCT), which captures the more subtle changes in the sound [27]. The Mel-spectrogram is enhanced by applying a Mel-scale filterbank to match the human auditory perception of sound [28]. Each feature was normalized and then stacked into the red channel (GFCC), green channel (spectrogram), and blue channel (Mel-spectrogram), forming a 224 × 224 RGB image. This representation enables the Vision Transformer to perform two-dimensional self-attention, effectively capturing both fine harmonic details and long-range temporal context.

Each cry waveform was transformed into three complementary 2D time-frequency representations: (i) a linear-scale spectrogram, (ii) a perceptually motivated Mel-spectrogram, and (iii) a Gammatone Frequency Cepstral Coefficient (GFCC) map. These three views were stacked as separate input channels to form a unified representation. This approach offers several advantages over using raw audio inputs. First, it enables the reuse of ImageNet-pretrained Vision Transformer weights, thereby reducing the number of randomly initialized parameters and significantly lowering computational and training resource requirements. Second, the combined views capture diverse acoustic characteristics (harmonic structure, perceptual loudness cues, and cepstral dynamics) resulting in a richer, more informative representation that enhances robustness to background noise. Third, prior research has demonstrated that spectrogram-based Vision Transformers can achieve competitive accuracy with lower GPU memory usage and faster training times compared to models that operate directly on raw audio signals [22].

This modification increased the ability of the model to identify different pathologies, as shown in Figures 2 and 3.

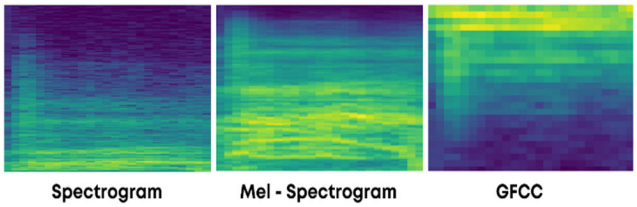

FIGURE 3. Examples of extracted audio features: Spectrogram, Mel-spectrogram, and GFCC from the audio signals.

From an acoustic perspective, stacking the spectrogram, Mel-spectrogram, and GFCC channels allows the model to capture locally correlated patterns that are expressed differently across feature types. While the spectrogram reflects the full linear frequency content, the Mel-spectrogram emphasizes perceptually important bands, and GFCCs model cochlear filtering to capture fine spectral cues. By aligning and stacking these features, shared acoustic structures, such as harmonics, formant trajectories, and onset patterns, are preserved. This integration enables the Vision Transformer to leverage complementary acoustic cues, enabling its ability to accurately detect pathological conditions.

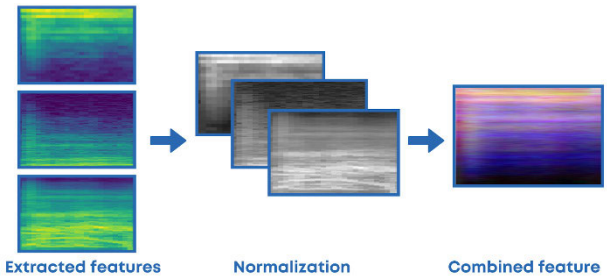

FIGURE 4. Normalization and combination of audio features into a single RGB image. Each feature serves a distinct role: the spectrogram captures energy patterns, the Mel-spectrogram reflects perceptual features, and GFCC represents compact speech features. Their combination provides richer input for the model.

In this study, the pathology–spectrum design bridge was made explicit: Sepsis and respiratory distress syndrome have been associated with irregular harmonics and attenuated high-frequency energy, which are captured by the spectrogram and Mel channels; jaundice and hyperbilirubinemia have been associated with reduced pitch variability and a narrower spectral spread, which are captured by the Mel and GFCC channels; and vomiting has been associated with time-localized broadband onsets and pitch breaks, which are captured by the spectrogram. These mappings were subsequently corroborated by LRP, LIME, and attention maps.

C. VISION TRANSFORMER

The selection of the ViT model is driven by its unique ability to capture both global and local contextual information through self-attention mechanisms, often outperforming traditional CNNs. ViT processes input by dividing images into non-overlapping patches, effectively addressing limitations related to regional interactions [20]. This technique works particularly well in identifying features of audio from spectrograms and results in greatly improved classification performance.

The implementation of the model included 12 transformer encoder layers, each incorporating multi-head self-attention and feed-forward networks. Input RGB images (224 × 224 pixels) were divided into 16 × 16 non-overlapping patches, which were then flattened and linearly embedded. Class tokens for each image patch were included, along with positional embeddings to preserve spatial relationships. These embeddings were processed through the transformer layers and a final classification head. To enhance performance on the limited dataset, transfer learning was applied using pre-trained ImageNet weights, enabling the model to achieve high accuracy and efficiency with minimal data.

The model parameters configuration is illustrated in Table 2. These values were determined through preliminary experiments and guided by the findings in Study [25], ensuring a balance between accuracy and training efficiency.

TABLE 2. Training parameters for the ViT model.

Parameter	Value
Cross-validation folds (k-fold)	5
Number of epochs	8
Learning Rate	8×10^{-5}
Weight Decay	6×10^{-3}
Batch Size	32
Number of Workers	8
Image Size	224 × 224 pixels

D. MODEL TRAINING AND EVALUATION

The ViT model was trained using the timm library, specifically with the pre-trained ViT-Base Patch16 224 weights for ImageNet. Hyperparameter tuning after initial experiments resulted in a batch size of 32, a learning rate of 7.9×10^{-5}, a weight decay of 6×10^{-3}, and an AdamW optimizer with OneCycleLR learning rate scheduler. Mixed precision training was used for cost-efficient computation. The model was trained for 8 epochs while training loss and validation accuracy were monitored. CrossEntropyLoss was implemented due to its efficient operation in multiclass complications.

To assess the performance of the proposed model, accuracy, precision, recall, and F1-score were computed and these metrics are described as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

E. CROSS VALIDATION

Validation using a five-fold cross-validation approach was rigorously applied to test the generalizability of the model. In each iteration, the data were divided into five subsets; four of them were used to train the model while the fifth would be used to validate it. All variations were taken into account such as 3-class, 4-class, 5-class, and 6-class problems to ensure thorough testing. The measures that were calculated within the folds were averaged to evaluate the overall accuracy and stability of the model.

F. PATHOLOGICAL INTERPRETATION

Before applying XAI methods, the known clinical and acoustic characteristics of each class in the dataset were thoroughly reviewed. Healthy infant cries typically exhibit a balanced energy distribution and stable harmonic structure. In contrast, cries from infants with Sepsis and RDS often show irregular harmonic spacing; notably, RDS cries tend to have reduced high-frequency energy due to respiratory compromise. Jaundice and hyperbilirubinemia are commonly associated with a monotonic pitch and a narrower spectral range, reflecting diminished neuromuscular control. Cries during vomiting are characterized by brief pitch breaks and fluctuating amplitude caused by abdominal strain. In our experiments, the LRP and LIME attention maps consistently highlighted frequency regions corresponding to these established acoustic patterns, demonstrating that the model's attention aligned well with known pathological markers.

Accordingly, clinical descriptors were linked to specific spectral correlates in the fused representation—stable harmonic stacks in healthy cries; aperiodicity and mid-band emphasis in sepsis; high-frequency attenuation in respiratory distress syndrome; bandwidth narrowing in jaundice and hyperbilirubinemia; and transient onsets in vomiting—and the same regions were highlighted by LRP, LIME, and attention maps, indicating that pathophysiologically consistent cues were used by the model.

G. EXPLAINABLE AI

In order to enhance explainability and transparency of the ViT model's actions, three XAI techniques were adopted: LIME, LRP, and transformer attention visualization. The explanations offered by LIME and LRP were tailored to the specific needs of the study and provided the versatility needed to better understand not only how the model works but also how and why decisions were made during the classification of infant cry audio features.

To understand the impact of specific audio features on singular predictions, LIME was applied. In this case, the ViT model was locally approximated with an interpretable one. LIME draws attention to the most relevant portions of the input—such as spectrograms, mel-spectrograms, and GFCCs—and classifies them according to their contribution to the classification outcome. This greatly assisted in understanding the attention mechanisms employed by the model, and, more importantly, the parts that played a critical role in the decision-making [29].

LRP identified the model's audial image inputs along with their respective relevance scores, tracing back estimation processes expertly. Assigning relevance via LRP enables output visualization through input image portions (constructed from audio features) across distinct layers to track back predictions made by the model. This technique enables analysis of the reasoning behind decisions made by the model, focusing attention on critical inputs instrumental to final outputs [23].

Using attention visualization within the transformer architecture, input spectrogram areas that the ViT model utilized during estimation processes were analyzed. Within the self-attention layers, attention maps synthesized by the model articulate the model's focus, shedding light on significant

frequency- and time-associated elements. Thus, this method offers insightful interpretation.

For the improvement of clinical interpretability, model transparency, and explainability, LIME and LRP were applied under the principles of XAI. For LIME, local surrogate models were designed for the machine learning model trained on the audio features to identify specific regions of interest, and for LRP, the model's predictions were dissected in layers to distribute a relevance score among the features that contributed to the outcome. The attention layers of the transformer further assisted in the visualization of critical sections of the spectrogram that were significant in forming the model's predictions. Collectively, these methods provided an understanding of the model's workings, reducing clinical distrust toward the use of sophisticated medical technologies and fostering their adoption in practice.

IV. EXPERIMENTAL RESULTS

A. PREFORMANCE EVALUATION

The Vision Transformer model was first tested for performance on an infant cry pathology classification task in different scenarios (3, 4, 5, and 6 class problems). Each scenario was optimized based on the maximum number of records available per class. For three classes, each class had 3000 records; four classes had 1158 records per class, five classes were balanced at 993 records per class, and six classes had 929 records per class.

Overall, the model performed impressively across all scenarios. For the three-class classification scenario, the model performed optimally, reporting an accuracy of 99.78% with precision, recall, and an F1-score of 100%. The model also performed optimally in the four-class scenario, achieving an accuracy of 99.42% with all other measures at 99%. The five-class scenario showed further improvement with 99.73% accuracy and precision, recall, and an F1-score of 1.00. In the six-class scenario, there was a slight drop in accuracy to 98.92%. However, this was accompanied by high precision, recall, and F1-score of 99% each, as shown in Table 3. Figure 5 contains confusion matrices alongside the loss and accuracy for each classification scenario for each epoch. From these results, it can be seen that the model was able to achieve high accuracy in relatively few epochs. For the three-class problem, the confusion matrix appears to illustrate perfect classification with almost no error made in distinguishing ''Sepsis'' from other classes. The loss curves also show an apparent convergence by the end of the 3rd epoch, and the accuracy curves provide clear indications of stabilization at near 100% starting from the 4th epoch. In the four-class problem, minor confusion is observed between classes such as ''Jaundice'' and other pathologies. The training and validation loss exhibit a peak in the second epoch and a sharp decrease later. At the same time, the accuracy seems to increase steadily, stabilizing at over 99% after the initial fluctuation.

For the five-class classification, some minor misclassifications occur between ''Vomiting'' and other classes, but these occurrences are very sparse, as shown in the confusion matrix. In this case, both the training and validation loss curves consistently declined and approached zero around the 8th epoch. The accuracy appears to have a steep increase and stabilizes around 99%, but this happened after the 4th epoch. In the six-class classification scenario, the minor confusion becomes slightly worse, especially between ''Hyperbilirubinemia'' and other classes. The training and validation losses dropped steadily and plateaued at low values by the end of training. Throughout the epochs, accuracy improved continuously until it plateaued at just under 99%, suggesting strong performance due to the added complexity.

The model was evaluated across four progressively challenging scenarios: starting with three-class (Healthy, Sepsis, and RDS), then adding Jaundice for a four-class setup, followed by the inclusion of Hyperbilirubinemia for five classes, and finally incorporating Vomiting to form a six-class classification.

TABLE 3. Performance results of VIT model without cross-validation.

Scenario	Accuracy	F1-score	Precision	Recall
3 classes	99.78%	100%	100%	100%
4 classes	99.42%	99%	99%	99%
5 classes	99.73%	100%	100%	100%
6 classes	98.92%	99%	99%	99%

B. CROSS VALIDATION RESULTS

To further evaluate the robustness and generalizability of the ViT model, a five-fold cross-validation was conducted on four balanced classification scenarios (three, four, five, and six classes). The results are shown in Table 4. In the three-class scenario (Healthy, RDS, Sepsis), the average accuracy in the five folds was 99.82%, while precision, recall, and F1-score maintained a perfect score of 100%. This result was consistent with minor fluctuations, indicating a very high level of reliability. In the four-class scenario (Healthy, RDS, Sepsis, Jaundice), the model achieved an average accuracy of 99.55% with class values for precision, recall, and F1-score also approaching a perfect score of 100%. Performance across folds was consistent and stable, with maximum accuracy of 100% achieved in fold 5. For the five-class classification scenario (Healthy, RDS, Sepsis, Jaundice, Hyperbilirubinemia), the average accuracy leveled off slightly to 99.05%, still achieving precision, recall, and F1-score of 99%. Variability between the folds was minimal, indicating strong performance even with increasing complexity. The six-class scenario (Healthy, RDS, Sepsis, Jaundice, Vomit, Hyperbilirubinemia) achieved an average accuracy of 99.34%, with precision, recall, and F1-score of 99%. Although the number of classes has increased, the results indicate a stable performance, strengthening the case in terms for model reliability in real-world testing.

From the cross-validation results, it is clear that the ViT model is able to achieve accuracy, precision, recall, and F1-scores regardless of the challenge posed by the scenario's

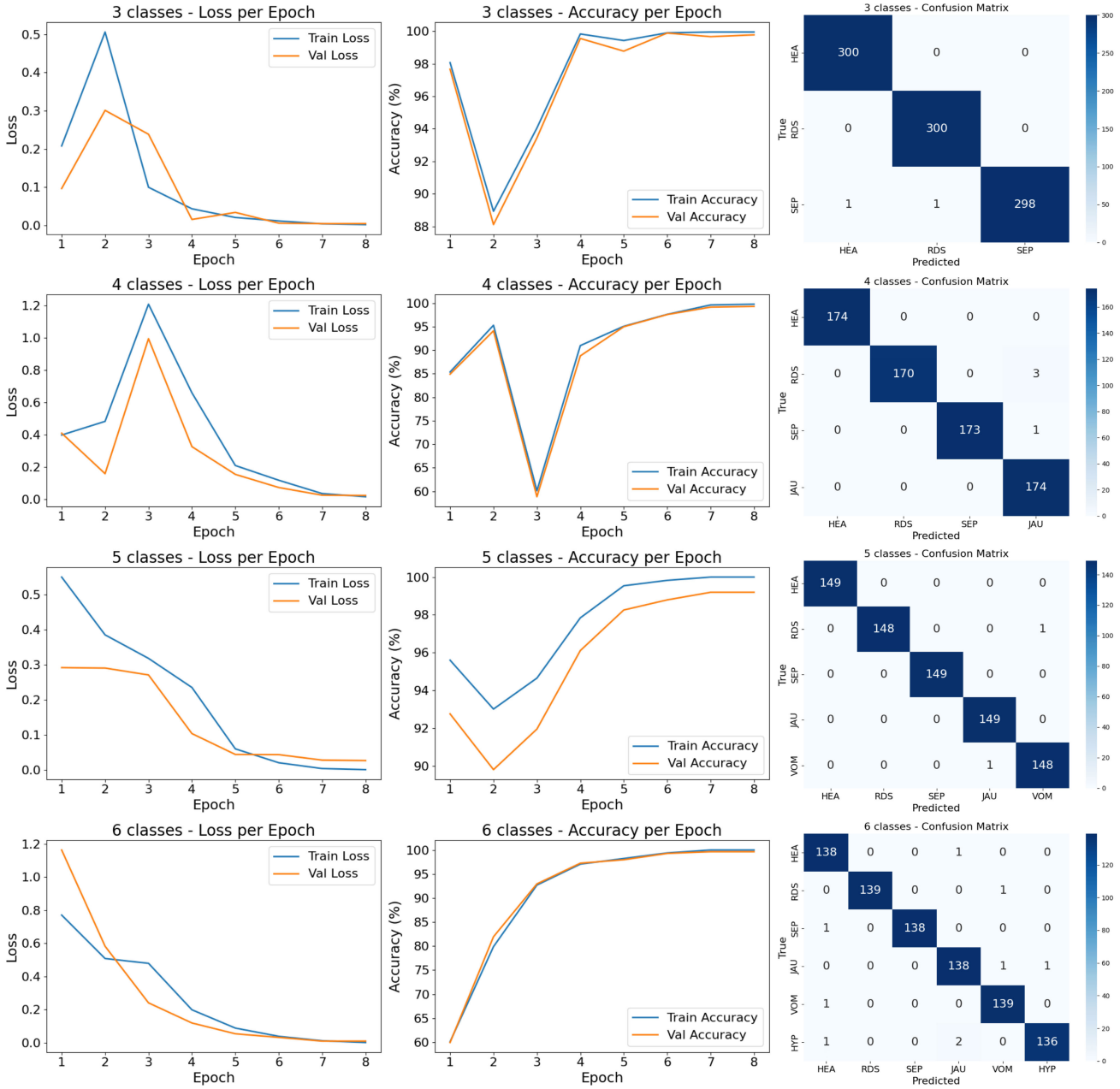

FIGURE 5. Training and validation loss, accuracy curves, and confusion matrices for the three, four, five, and six-class classification scenarios.

TABLE 4. Cross-validation results across different classification scenarios across 5 folds.

Scenario	AVG. Accuracy	AVG. F1-score	AVG. Precision	AVG. Recall
3 classes	99.82%	100%	100%	100%
4 classes	99.55%	100%	100%	100%
5 classes	99.05%	99%	99%	99%
6 classes	99.34%	99%	99%	99%

complexity. The strength demonstrated by cross-validation also reflects the appropriateness of the model for clinical use in classifying infant cry pathologies.

C. EXPLAINABLE AI ANALYSIS

XAI methods such as LIME, LRP, and attention-based models were utilized to explain the decision processes of the ViT model. Aggregate results for each method across all classes are displayed in Figures 6, 7, and 8.

According to the LRP analysis (Figure 7), some spectral regions had a high power for class separation. For example, healthy cry signals activated upper-frequency ranges, suggesting stable and distinctive low-level patterns associated with non-pathological crying. In contrast, classes like RDS and Sepsis showed strong activation across many frequency bands, indicating of important discriminative low- and mid-frequency pathologic ranges.

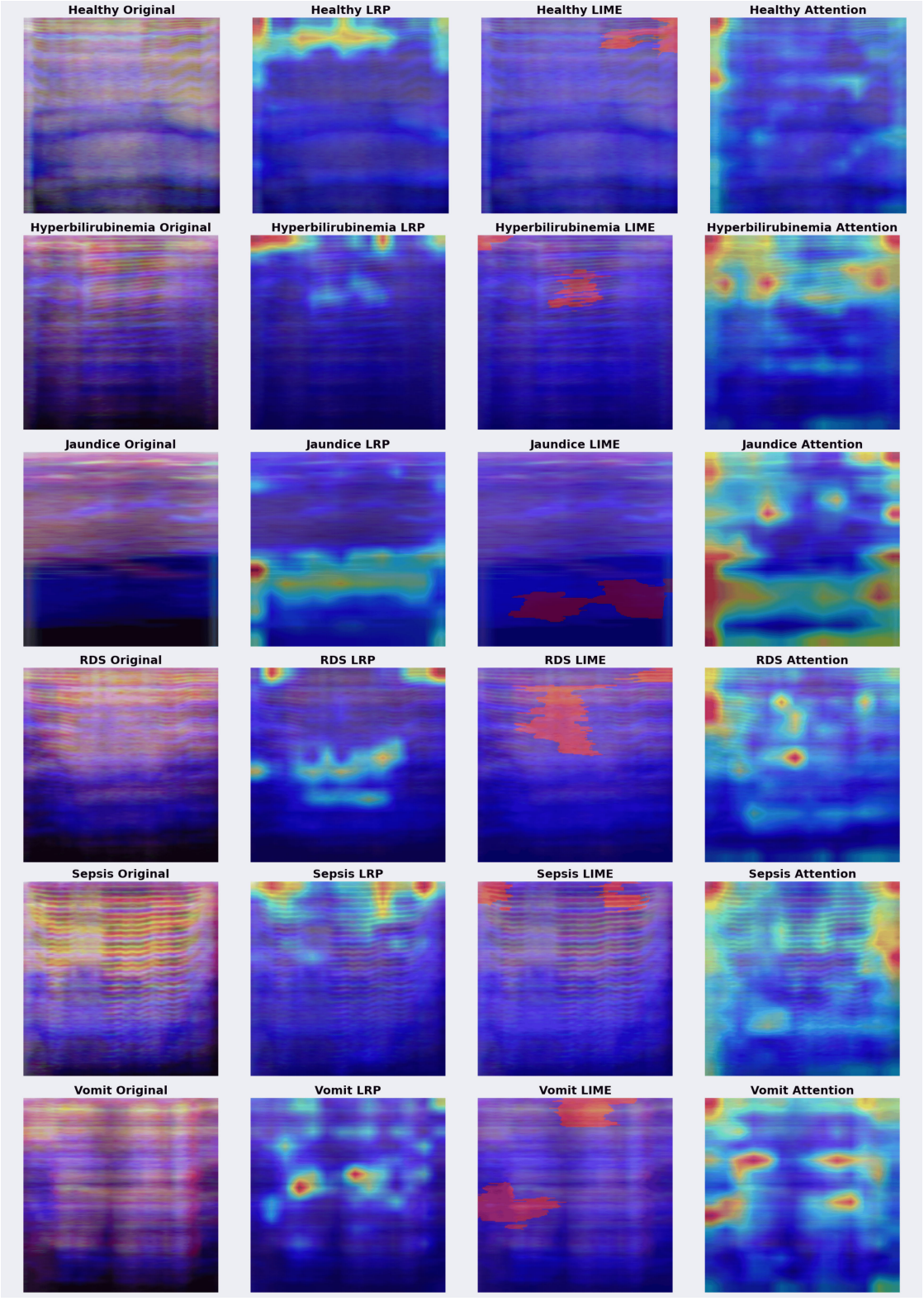

FIGURE 6. Explainable AI for Combined Audio Features: GFCC, Spectrogram, and Mel-Spectrogram Using LRP, LIME, and Attention Mechanisms.

LIME diagrams, as shown in Figure 8, provide class-level explanations and support this analysis by showing which parts of each spectrogram affected the model decision most. For example, LIME provided the most relevant

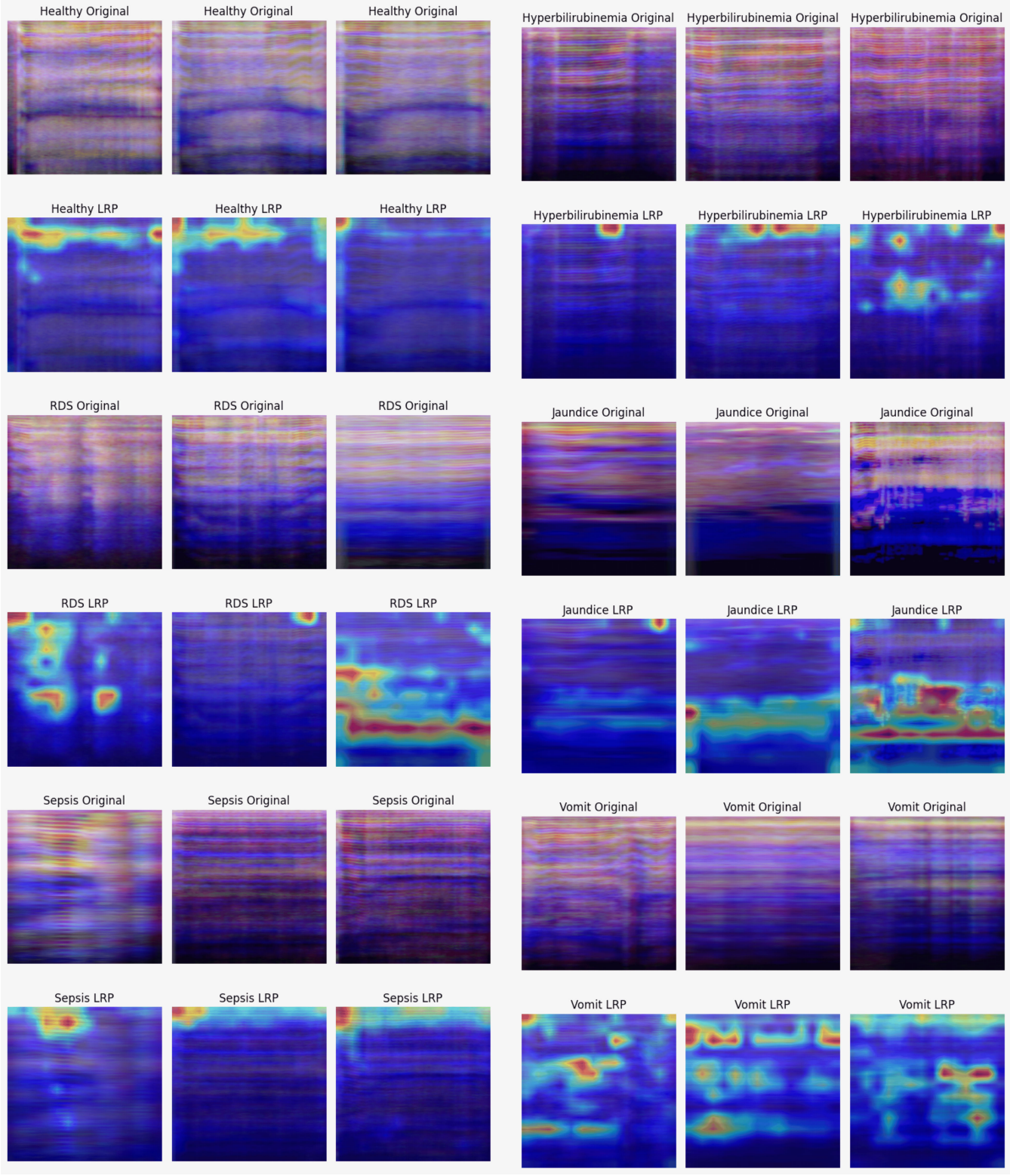

FIGURE 7. Visualization of ViT model interpretations using LRP.

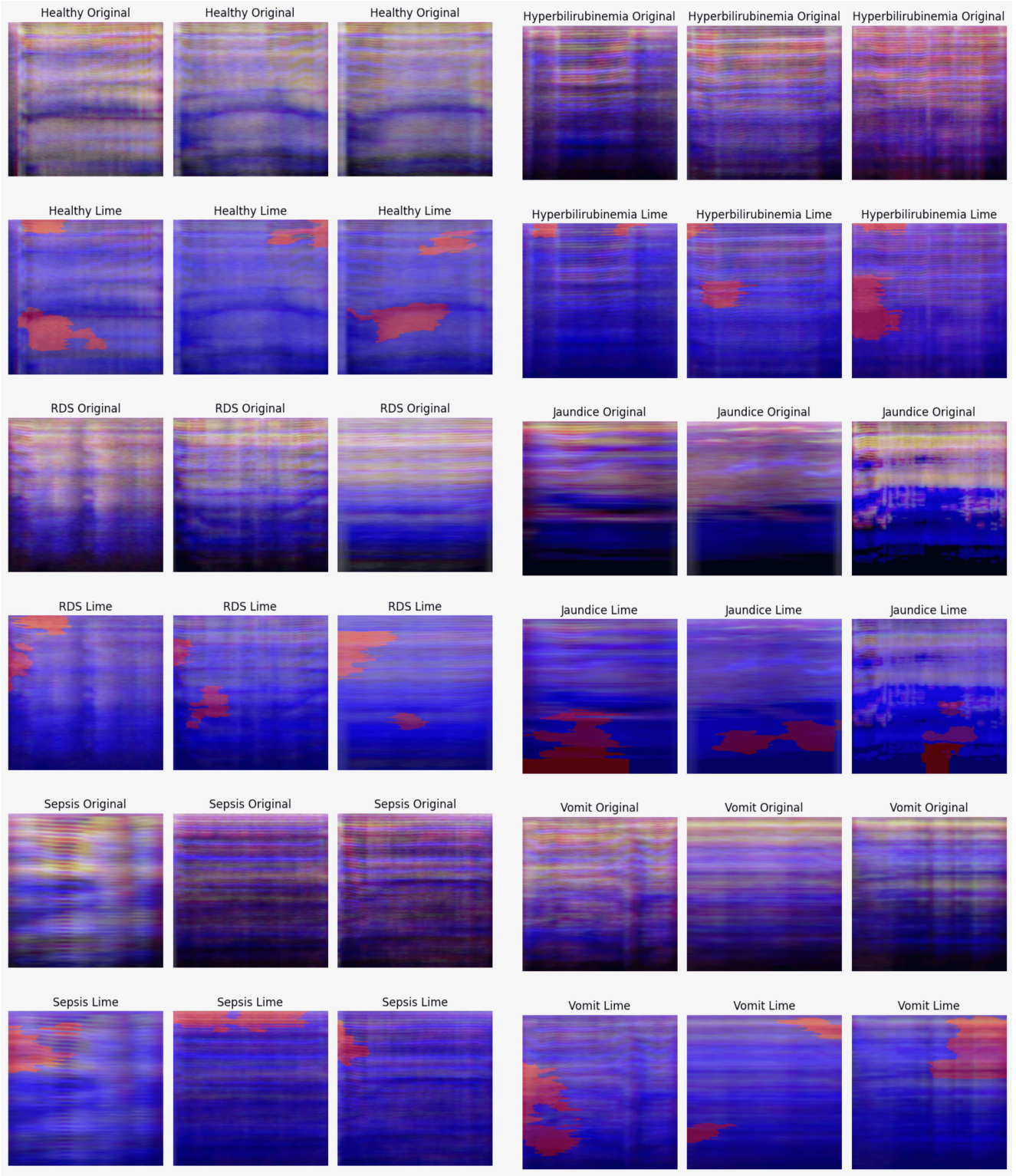

FIGURE 8. Visualization of ViT model interpretations using LIME.

temporal-frequency region for Jaundice and Hyperbilirubinemia, which was a natural inference given the high classification accuracy achieved.

On the other hand, attention-based models, shown in Figure 6, also provided a global view of the effects, with the output of the Attention LSTM revealing more broadly associated features and helping to make correct predictions, showing disparity for all audio inputs of Sepsis, RDS, and revealing powerful spectral structures unique to the above pathologies.

TABLE 5. Comparison of classification results of the proposed study with previous studies.

Study	Classes	Audio Features	ML Algorithm	Accuracy	Precision	Recall	F1-score (%)
Model [11]	2	GFCCs, HR	MLP	95.92%	95%	95%	95%
Model [19]	3	GFCCs, HR, spectrogram	Fusion CNN	97.5%	97.51%	97.53%	97.52%
Model [18]	3	Spectrogram	SVM + CNN	92.5%	88.8%	89.3%	88.9%
Model [30]	4	LFCCs	XGBoost	92%	-	-	92.3%
Model [22]	3	Spectrogram	AST	98.69%	98.73%	98.71%	98.71%
Model [25]	3	GFCC: 96.33% Spectrogram: 93.17% Mel Spectrogram: 94.83%	Vision Transformer and XAI	GFCC: 96.33 Spectrogram: 93.17% Mel Spectrogram: 94.83%	GFCC: 96% Spectrogram: 93% Mel Spectrogram: 95%	GFCC: 96% Spectrogram: 93% Mel Spectrogram: 95%	GFCC: 96%, Spectrogram: 93%, Mel Spectrogram: 95%
Model [31]	3	wav2vec 2.0, WavLM, HuBERT, MFCCs, Gammatone Cepstral Coefficients, Spectral Features	SSL models, Transformer-based encoders, Softmax classifier, Adam optimizer, Annealing & Linear Learning Rate strategies	wav2vec 2.0: 89.76, WavLM Base+: 88.97%, HuBERT: 88.33%	wav2vec 2.0 (Healthy: 91.21%, Sepsis: 90.39%, RDS: 87.76%)	wav2vec 2.0 (Healthy: 91.43%, Sepsis: 87.38%, RDS: 90.48%)	wav2vec 2.0 (Healthy: 91.32%, Sepsis: 88.86%, RDS: 89.10%)
Proposed model	**6**	**Spectrogram, GFCC and Mel-Spectrogram combined**	**Vision transformer and XAI**	**99%**	**99%**	**99%**	**99%**

In summary, the integration of LRP with LIME and the infusion of attentional mechanisms allowed for a complete understanding of how the ViT model works. These visual elucidations substantiated the model's dependencies on relevant clinical audio features, further enhancing its interpretability and credibility in medical diagnostics.

V. DISCUSSION

By incorporating a multi-channel approach that combines three key audio features- Spectrogram, Mel-spectrogram, and GFCCs- into a single RGB image, this system marks a significant advancement over previous infant cry classification methods. As demonstrated in Table 5, earlier studies consistently suffered in improving accuracy, precision, recall, and F1-score, and none successfully classified more than six pathological classes, which is an achievement realized in this study. A superior efficiency/accuracy trade-off is obtained through the fusion of spectrogram, Mel-spectrogram and GFCC representations when compared with the use of any single view or raw audio alone.

In study [11], the authors employed GFCC features in combination with heart rate (HR) data, applying them to a Multi-Layer Perceptron (MLP) model. This approach achieved an accuracy of 95.92%, though it was limited to only two classes. While the integration of physiological data (HR) contributed to a higher level of accuracy, however the model lacked generalizability. In contrast, the proposed approach in this study demonstrated the ability to handle larger number of classes with even higher accuracy, offering a more robust and scalable solution to the problem of infant cry-based diagnostics.

In study [19], the authors applied a fusion deep learning approach that integrated GFCC, HR, and spectrogram features for three-class classification. While higher accuracy was reported (97.50%), the added data modalities (audio and HR) introduced additional difficulties in implementation and clinical use. The proposed model showed better performance with added ease of implementation, using only audio features processed straightforwardly directly by a ViT.

With a hybrid of CNN and SVM, with spectrogram classifier features were used in [18], yielding an accuracy of 92.50% for three classes. This combination provided additional complexity and relatively lower performance than the streamlined version of the ViT model presented here. Again, this model showed superior accuracy along with ease of deployment and training.

In study [30], the XGBoost algorithm was used in combination with LFCC features for four-class classification, achieving an accuracy of 92%. Despite utilizing this advanced boosting algorithm, the performance did not surpass that of the current method. In contrast, the proposed ViT-based architecture, combined with a multi-feature approach, demonstrated superior accuracy and reliability across a broader set of classes, delivering performance that remains unmatched

by previous methods. Study [22] utilized an AST model, which processed audio signals in their raw form through a succession of computational steps without prior adjustment or feature extraction, achieving an accuracy of 98.69% for three classes (Healthy, Sepsis, and RDS). While the AST model performed remarkably well with the raw audio, it was extremely overengineered and required too much optimization. On the other hand, the current study used a ViT model that worked with processed spectrogram-based features, greatly simplifying the training required and the computations needed. This approach improved the accuracy, as well as the efficiency, scalability, and resource requirements of the model.

In study [25], the authors implemented a ViT model for audio classification; however, the audio features were evaluated separately, achieving the highest accuracy of 96.33% with the GFCC-based model. In contrast, the current research employed a synergistic approach by integrating all three audio features (GFCC, Mel-spectrogram, and spectrogram) within the ViT model. This fusion led to unprecedented performance, with a classification accuracy reaching 99%. Additionally, the study utilized exhaustive cross-validation, ensuring that each feature channel was thoroughly evaluated, thereby enhancing the model's generalization capabilities beyond the isolated feature approach used in study [25].

Study [31] had several models utilizing more advanced self-supervised learning techniques like wav2vec 2.0, WavLM, and HuBERT with convolutional transformers as encoders, which yielded accuracies of 89.76%. While these methods were able to extract deep audio features using advanced SSL techniques, the approach proposed in this study achieved higher accuracy results in terms of precision, recall, and F1 score using a combination of simpler robust audio features and a ViT architecture. These results significantly outperformed the other techniques.

Through the use of explainable AI techniques such as LRP, LIME, and attention mechanisms, XAI provided complete transparency that set this research apart from others and justified its use. Figures 6–8 provide detailed visualizations of the model's explanations. In particular, the LRP heat maps highlight low- to mid-frequency bands when classifying Sepsis and RDS, while relevance for Healthy cries is primarily concentrated in higher-frequency regions. Similarly, the LIME overlays isolate these discriminative areas, reinforcing the importance of those spectral cues. Attention roll-outs further demonstrate that the Vision Transformer focuses on specific time-frequency regions where pathological cries differ from healthy patterns, offering insight into the model's high precision. These powerful interpretability methods helped uncover the constituent and most critical audio features on which the model relied on for its predictions, enhancing both trust in and understanding of the model within a clinical setting.

VI. CONCLUSION AND FUTURE WORK

A novel framework for infant-cry classification has been introduced, leveraging an RGB image representation that integrates Spectrogram, Mel-spectrogram, and GFCCs features. Using a ViT model, the proposed model achieved significantly higher classification accuracy compared to previous recent CNN-based approaches due to ViT's ability to aggregate contextual information at a global scale. The tri-modal representation harnesses complementary acoustic cues, while the self-attention mechanism of the ViT captures long-range temporal patterns, resulting in a classification accuracy of 99 % across six neonatal conditions. This demonstrates not only the performance advantage of the proposed model, but also the effectiveness of multi-feature representations in audio diagnostics.

The integration of XAI techniques, such as LRP, LIME, and attention, enhances model interpretability. These techniques provide a pixel-level visual explanation that offer transparency into the model's decision-making process, thereby fostering greater trust in automated neonatal diagnostic systems in particular.

Future work will focus on three key directions: (i) developing lightweight GRU-augmented or Tiny-ViT variants appropriate for real-time embedded deployment; (ii) constructing a larger multimodal dataset that combines cry acoustics with respiration signals, facial expressions, and realistic background noise; and (iii) exploring self-supervised and semi-supervised learning strategies to improve model robustness under varied clinical recording conditions.

ACKNOWLEDGMENT

(Ahmad Hasasneh and Sari Masri contributed equally to this work.)

REFERENCES

[1] W. B. UNICEF WHO and United Nations. (2024). *Levels and Trends in Child Mortality: Report 2024*. UNICEF. [Online]. Available: https://data.unicef.org/resources/levels-and-trends-in-child-mortality-2024/

[2] World Health Organization. (2024). *Newborns: Reducing Mortality-Fact Sheet*. World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality

[3] Merck Manual Professional Version. (2024). *Neonatal Sepsis*. Merck & Co., Inc. [Online]. Available: https://www.merckmanuals.com/professional/pediatrics/infections-in-neonates/neonatal-sepsis

[4] B. D. Kamath, E. R. MacGuire, E. M. McClure, R. L. Goldenberg, and A. H. Jobe, "Neonatal mortality from respiratory distress syndrome: Lessons for low-resource countries," *Pediatrics*, vol. 127, no. 6, pp. 1139–1146, Jun. 2011, doi: 10.1542/peds.2010-3212.

[5] B. O. Olusanya, M. Kaplan, and T. W. R. Hansen, "Neonatal hyperbilirubinaemia: A global perspective," *Lancet Child Adolescent Health*, vol. 2, no. 8, pp. 610–620, Aug. 2018, doi: 10.1016/s2352-4642(18)30139-1.

[6] B. Ansong-Assoku, S. D. Shah, M. Adnan, and P. A. Ankola, *Neonatal Jaundice*. Treasure Island, FL, USA: StatPearls Publishing, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK532930/

[7] M. K. Ghiam, A. Langerman, Z. Sargi, and S. Rohde, "Head and neck cancer patients: Rates, reasons, and risk factors for 30-day unplanned readmission," *Otolaryngol.–Head Neck Surg.*, vol. 159, no. 1, pp. 149–157, Jul. 2018, doi: 10.1177/0194599818776633.

[8] E. E. Turhan, T. Gürsoy, and F. Ovali, "Factors which affect mortality in neonatal sepsis," *Türk Pediatri Arşivi*, vol. 50, no. 3, pp. 170–175, Sep. 2015, doi: 10.5152/turkpediatriars.2015.2627.

[9] K. Michelsson, P. Sirviö, and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," *Acta Paediatrica*, vol. 66, no. 5, pp. 611–616, Sep. 1977, doi: 10.1111/j.1651-2227.1977.tb07956.x.

[10] Y. Kheddache and C. Tadj, "Characterization of pathologic cries of newborns based on fundamental frequency estimation," *Engineering*, vol. 5, no. 10, pp. 272–276, 2013, doi: 10.4236/eng.2013.510b057.

[11] Z. Khalilzad, A. Hasasneh, and C. Tadj, "Newborn cry-based diagnostic system to distinguish between sepsis and respiratory distress syndrome using combined acoustic features," *Diagnostics*, vol. 12, no. 11, p. 2802, Nov. 2022, doi: 10.3390/diagnostics12112802.

[12] A. T. Patil, A. Kachhi, and H. A. Patil, "Subband teager energy representations for infant cry analysis and classification," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2022, pp. 1313–1317, doi: 10.23919/EUSIPCO55093.2022.9909974.

[13] M. Hariharan, L. S. Chee, and S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network," *J. Med. Syst.*, vol. 36, no. 3, pp. 1309–1315, Jun. 2012, doi: 10.1007/s10916-010-9591-z.

[14] P. I. Rani, P. P. Kumar, V. M. Immanuel, P. Tharun, and P. Rajesh, "Baby cry classification using machine learning," *Int. J. Innov. Sci. Res. Technol.*, vol. 7, no. 2, pp. 561–566, 2022. [Online]. Available: https://ijisrt.com/assets/upload/files/IJISRT22MAR645.pdf

[15] F. Salehian Matikolaie, Y. Kheddache, and C. Tadj, "Automated newborn cry diagnostic system using machine learning approach," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103434, doi: 10.1016/j.bspc.2021.103434.

[16] G. Z. Felipe, R. L. Aguiar, Y. M. G. Costa, C. N. Silla, S. Brahnam, L. Nanni, and S. McMurtrey, "Identification of Infants' cry motivation using spectrograms," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jun. 2019, pp. 181–186, doi: 10.1109/IWSSIP.2019.8787318.

[17] C.-Y. Chang and J.-J. Li, "Application of deep learning for recognizing infant cries," in *Proc. IEEE Int. Conf. Consum. Electronics-Taiwan (ICCE-TW)*, May 2016, pp. 1–2, doi: 10.1109/ICCE-TW.2016.7520947.

[18] K. Ashwini, P. M. D. R. Vincent, K. Srinivasan, and C.-Y. Chang, "Deep learning assisted neonatal cry classification via support vector machine models," *Frontiers Public Health*, vol. 9, Jun. 2021, Art. no. 670352, doi: 10.3389/fpubh.2021.670352.

[19] Y. Zayed, A. Hasasneh, and C. Tadj, "Infant cry signal diagnostic system using deep learning and fused features," *Diagnostics*, vol. 13, no. 12, p. 2107, Jun. 2023, doi: 10.3390/diagnostics13122107.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[21] P. Verma and J. Berger, "Audio transformers," 2021, *arXiv:2105.00335*.

[22] M. Tami, S. Masri, A. Hasasneh, and C. Tadj, "Transformer-based approach to pathology diagnosis using audio spectrogram," *Information*, vol. 15, no. 5, p. 253, Apr. 2024, doi: 10.3390/info15050253.

[23] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2022, pp. 193–209. [Online]. Available: https://doi.org10.1007/978-3-030-28954-6_10

[24] P. Komorowski, H. Baniecki, and P. Biecek, "Towards evaluating explanations of vision transformers for medical imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3726–3732, doi: 10.1109/CVPRW59228.2023.00383.

[25] S. Masri, A. Hasasneh, M. Tami, and C. Tadj, "Exploring the impact of image-based audio representations in classification tasks using vision transformers and explainable AI techniques," *Information*, vol. 15, no. 12, p. 751, Nov. 2024, doi: 10.3390/info15120751.

[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020, doi: 10.1109/TASLP.2020.3030497.

[27] B. Ayoub, K. Jamal, and Z. Arsalane, "Gammatone frequency cepstral coefficients for speaker identification over VoIP networks," in *Proc. Int. Conf. Inf. Technol. Organizations Develop. (IT4OD)*, Mar. 2016, pp. 1–5, doi: 10.1109/IT4OD.2016.7479293.

[28] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396, doi: 10.1109/ICASSP.2017.7952585.

[29] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*. San Diego, CA, USA: Association for Computational Linguistics, 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.

[30] V. R. Joshi, K. Srinivasan, P. M. D. R. Vincent, V. Rajinikanth, and C.-Y. Chang, "A multistage heterogeneous stacking ensemble model for augmented infant cry classification," *Frontiers Public Health*, vol. 10, 2022, pp. 1–9, doi: 10.3389/fpubh.2022.819865.

[31] S. V. Shayegh and C. Tadj, "Deep audio features and self-supervised learning for early diagnosis of neonatal diseases: Sepsis and respiratory distress syndrome classification from infant cry signals," *Electronics*, vol. 14, no. 2, p. 248, Jan. 2025, doi: 10.3390/electronics14020248.

AHMAD HASASNEH received the B.Sc. degree in computer systems engineering (CSE) from Palestine Polytechnic University (PPU), in 2005, the M.Sc. degree in computer graphics and programming from the University of Hull, U.K., in 2006, and the Ph.D. degree (Hons.) in computer science, majoring in machine learning, in 2012. In 2009, he was offered a full scholarship to join the Ph.D. program at Paris University. He was a full-time Lecturer with the Department of Computer Science, Hebron University, for more than three years. He re-joined the Computer Department, Hebron University, as an Assistant Professor, for two years. From 2014 to 2015, he was the Head of the Computer Department, Palestine Technical University. From 2015 to 2021, he was an Assistant Professor and the Dean of the Engineering and Information Technology College, Palestine Ahliya University (PAU). Since 2021, he has been with the Faculty of Graduate Studies, Arab American University (AAU), and currently, he is heading the Data Science Department. He has several international joint research projects, particularly within Palestinian German Science Bridge and Palestinian Quebec Science Bridge. Moreover, he has published over 15 international journals and conferences, and his research interests include machine learning, deep learning, robotics, feature extraction, recognition, robot localization, neurosciences, image processing, and segmentation.

SARI MASRI received the B.Sc. degree in computer information systems from Bethlehem University, Palestine. He is currently pursuing the M.Sc. degree in artificial intelligence with Arab American University, Palestine. He has worked on projects involving vision transformers, generative AI models, and retrieval-augmented generation (RAG) techniques. In addition to his academic work, he has extensive industry experience as a Senior Backend Developer, specializing in cloud architectures, microservices, and serverless solutions on AWS. He has co-authored research papers on audio classification and traffic conflict detection using LLMs, with publications in recognized journals and conferences. His research interests include deep learning, computer vision, large language models (LLMs), and explainable artificial intelligence (XAI), with applications in healthcare diagnostics, traffic management, and multimodal AI systems. His broader research interests include vision-based recognition, ethical AI practices, robotics, and multimodal learning.

CHAKIB TADJ received the Ph.D. degree in signal and image processing from ENST Paris, in 1995. He is currently a Professor with the École de Technologie Supérieure (ETS), University of Quebec, Montreal, Canada. His research interests include signal processing, speech recognition, pervasive computing, and multimodal systems.