information

MDPI

Article

Exploring the Impact of Image-Based Audio Representations in Classification Tasks Using Vision Transformers and Explainable AI Techniques

Sari Masri [1], Ahmad Hasasneh [1,*], Mohammad Tami [1] and Chakib Tadj [2]

[1] Department of Natural, Engineering and Technology Sciences, Faculty of Graduate Studies, Arab American University, Ramallah P.O. Box 240, Palestine; s.masri3@student.aaup.edu (S.M.); m.abutami@student.aaup.edu (M.T.)

[2] Department of Electrical Engineering, École de Technologie Supérieur, Université du Québec, Montreal, QC H3C 1K3, Canada; chakib.tadj@etsmtl.ca

* Correspondence: ahmad.hasasneh@aaup.edu

Abstract: An important hurdle in medical diagnostics is the high-quality and interpretable classification of audio signals. In this study, we present an image-based representation of infant crying audio files to predict abnormal infant cries using a vision transformer and also show significant improvements in the performance and interpretability of this computer-aided tool. The use of advanced feature extraction techniques such as Gammatone Frequency Cepstral Coefficients (GFCCs) resulted in a classification accuracy of 96.33%. For other features (spectrogram and mel-spectrogram), the performance was very similar, with an accuracy of 93.17% for the spectrogram and 94.83% accuracy for the mel-spectrogram. We used our vision transformer (ViT) model, which is less complex but more effective than the proposed audio spectrogram transformer (AST). We incorporated explainable AI (XAI) techniques such as Layer-wise Relevance Propagation (LRP), Local Interpretable Model-agnostic Explanations (LIME), and attention mechanisms to ensure transparency and reliability in decision-making, which helped us understand the why of model predictions. The accuracy of detection was higher than previously reported and the results were easy to interpret, demonstrating that this work can potentially serve as a new benchmark for audio classification tasks, especially in medical diagnostics, and providing better prospects for an imminent future of trustworthy AI-based healthcare solutions.

Keywords: vision transformers (ViTs); infant cry classification; audio signals; image-based representations; gammatone frequency cepstral coefficients (GFCCs); spectrogram; mel-spectrogram; explainable AI (XAI); layer-wise relevance propagation (LRP); local interpretable model-agnostic explanations (LIME); audio feature; medical diagnostics; healthcare AI

Citation: Masri, S.; Hasasneh, A.; Tami, M.; Tadj, C. Exploring the Impact of Image-Based Audio Representations in Classification Tasks Using Vision Transformers and Explainable AI Techniques. *Information* **2024**, *15*, 751. https://doi.org/10.3390/info15120751

Academic Editor: Danilo Avola

Received: 6 October 2024
Revised: 14 November 2024
Accepted: 21 November 2024
Published: 25 November 2024

1. Introduction

The exploration of image-based audio representations for classification tasks takes advantage of significant advancements in ViT [1] and explainable AI (XAI) techniques, offering a new approach to combining audio and visual data for improved classification accuracy. Unlike classical machine learning (ML) or deep learning (DL) methods, such as Support Vector Machines (SVMs), which often require comprehensive feature engineering and may struggle with complex, high-dimensional audio data, ViTs excel due to their ability to model global context through self-attention mechanisms [2]. By converting audio signals into audio features, which can be treated as images, ViTs, originally designed for image classification, have shown excellent performance in various domains, including audio classification [3]. This approach enhances accuracy and allows for greater explainability and the possibility of integrating ViTs with other methods to boost performance further.

Vision transformers have revolutionized the field of image classification by employing a transformer-based architecture that processes image patches as sequences, akin to words in natural language processing. The self-attention mechanism of ViTs allows them to capture long-range dependencies and hierarchical representations more effectively than traditional Convolutional Neural Networks (CNNs). This is particularly advantageous for classification tasks where understanding the global context is crucial. Recent studies have demonstrated the superiority of ViTs over CNNs in both accuracy and generalization, particularly when trained on large datasets such as ImageNet and CIFAR-10 [4,5].

The rest of this paper is organized as follows: the Literature Review reflects the existing work in image-based audio representations, vision transformers, and explainable AI-driven strategies that can help improve model interpretability. The dataset is described in terms of the origins, processing pipeline, and audio data representation and visualization. The Materials and Methods section provides the structure of the vision transformers, training details (where available), and the means by which explainable AI concepts are applied. The Results and Discussion sections present model performance metrics, compare them with traditional approaches, and provide the insights obtained using explainability tools. The results and contributions of this study are followed by a Future Work section, which suggests directions for further research in this domain. The Conclusion, which is followed by Future Work, summarizes the results and outlines the contributions of this study, as well as suggesting some directions for further research in this domain.

2. Literature Review

Transforming audio signals into visual representations, such as spectrograms, allows for the application of image-based DL techniques to audio classification tasks. Spectrograms provide a time–frequency representation of audio signals, making them suitable for processing with image-based models like ViTs. This methodology benefits from the advanced feature extraction abilities of ViTs, which can identify complex patterns within the spectrograms, enhancing classification performance in tasks such as speech recognition and environmental sound classification [6,7]. Integrating XAI techniques with ViTs is important for enhancing the interpretability of these models. Techniques such as LRP and attention visualization have provided insights into model decisions by attributing importance to input features. These methods are important for applications in critical domains such as medical imaging, where understanding the rationale behind model predictions is vital for trust and reliability [8,9]. Recent studies have focused on evaluating the explanations provided by ViTs, particularly in medical imaging tasks. For instance, ref. [10] investigated the performance of various interpretation methods on a ViT model applied to chest X-ray classification. They introduced metrics for evaluating the faithfulness, sensitivity, and complexity of ViT explanations, finding that LRP [11] outperforms other techniques regarding accuracy and reliability.

In a related study, ref. [12] explored the underlying mechanisms of multi-head self-attentions (MSAs) in vision transformers. Their research showed that MSAs improve the accuracy and generalization by flattening the loss landscape. This improvement is mainly attributed to data specificity rather than long-range dependency, highlighting the importance of dataset quality in training ViTs. They also proposed a model called AlterNet, which replaces the convolutional blocks at the end of a stage with MSA blocks and outperforms traditional CNNs in both large and small data settings. Moreover, ref. [13] provided a foundational study on applying vision transformers to large-scale image recognition tasks. Their work showed that ViTs, when trained on large datasets like ImageNet, achieve state-of-the-art performance by treating image patches as sequences and leveraging transformer architectures from natural language processing (NLP). This method allows ViTs to capture comprehensive features and dependencies within image data, leading to superior classification results compared to conventional CNNs. Also, ref. [6] extended the use of transformers to audio data, introducing Audio Transformers designed for large-scale audio understanding. By converting audio signals into spectrograms and processing them as images, the study showed significant improvements in audio classification. This approach has opened new possibilities for applying image-based techniques to audio-related problems and showcases the versatility of transformer architectures.

Recent research, including [10], has shown that while ViTs can make highly accurate predictions, there are still challenges with how well these predictions can be explained, especially in fields like healthcare. The study pointed out that current methods do not always provide clear and understandable explanations for the decisions made by ViTs, which is a problem when these models are used in critical areas where trust and transparency are important. Our proposed model uses ViTs combined with advanced XAI techniques to improve both prediction accuracy and the transparency of explanations. By using methods like LRP and LIME, we aim to make the model's decisions more transparent and easier to understand. This approach is designed to make ViTs more suitable for use in important fields like healthcare.

This paper investigates how different image-based audio representations could impact classification performance and add new understanding to the current state-of-the-art. The models can learn from different types of representations, such as spectral, mel-spectral, Gammatone Frequency Cepstral Coefficients (GFCCs), mel-frequency cepstral coefficients (MFCCs), and Rasta-PLP, by currently published implementations. Also, temporal features including tempograms, chroma, and raw waveform can feasibly be trained with our model architecture. This study will utilize more robust XAI techniques to improve the interpretability of model predictions and thereby increase the reliability for practical applications. This study will be a valuable contribution to the ease of optimizing ViTs for audio classification tasks and building more explainable AI.

3. Materials and Methods

In this study, a methodical approach was used to obtain a systematic and accurate result, which is characterized by five general stages as shown in Figure 1. Crying infant audio data were first gathered and segmented. The audio signals were converted into feature images using various feature extraction techniques such as spectrogram, mel-spectrogram, and MFCCs to transform the audio signals into feature images. These were effectively used to cover the training and evaluation of a basic vision transformer model. The feature list was evaluated using the same model architecture. The final phase was to bring explainable AI techniques LIME and LRP to interpret and clarify model decisions, thereby increasing the transparency and explanation of the results.

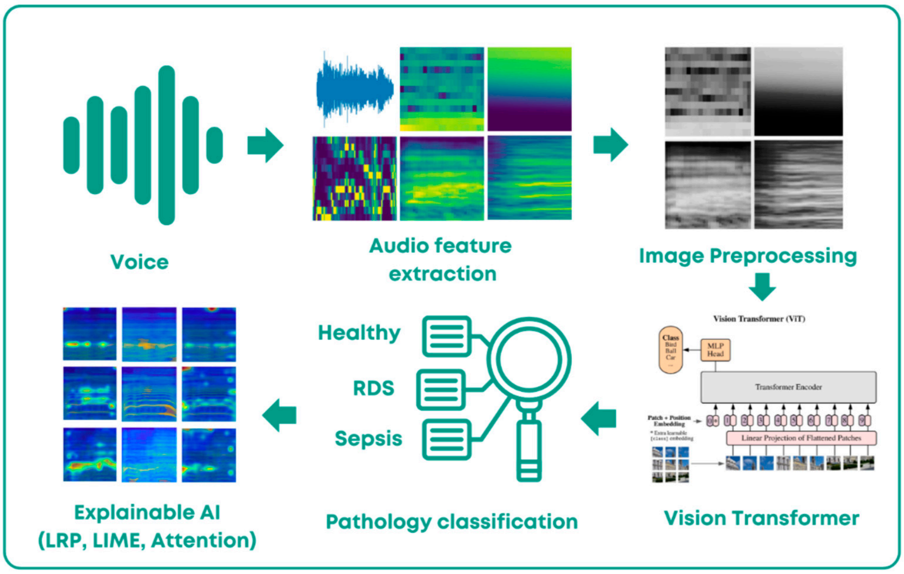

Figure 1. The workflow of the proposed vision transformer model incorporating explainable AI techniques.

3.1. Dataset Description

The dataset was collected at Al-Raee and Al-Sahel hospitals in Lebanon and Saint-Justine Children's Hospital in Montreal and has been used in several related research studies [8,14,15]. The dataset consists of crying newborn infants between 1 and 53 days of age from different locations and backgrounds. As shown in Table 1, it consists of 181 recordings from 83 newborns for the healthy class, 102 recordings from 33 newborns for the Respiratory Distress Syndrome (RDS) class, and 53 recordings from 17 newborns for the sepsis class. Each recording lasts an average of 90 s and was recorded five times for each newborn. The recordings were then segmented and labeled according to the model of previous researchers [3,14,16]. This labeling was performed using the WaveSurfer software (1.8.8). Recordings were made using a 2-channel Olympus digital recorder with 16-bit resolution and a sampling rate of 44,100 Hz, positioned 10 to 30 cm away from the infants [3]. The dataset includes a gender distribution of approximately 65% male (11 infants) and 35% female (6 infants) [14]. The final dataset comprises 2000 WAV records for each pathology, as shown in Figure 2.

Table 1. Dataset description.

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

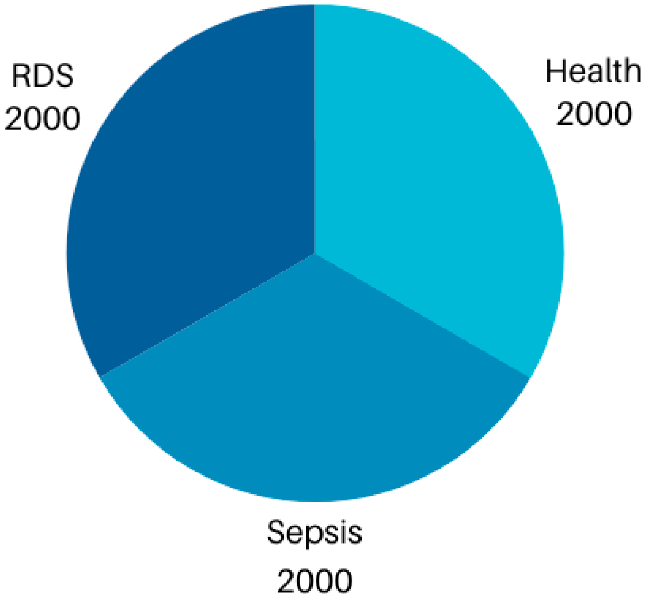

Figure 2. Distribution of the three classes of samples.

3.2. Feature Extraction

In this research, six different audio feature images were extracted from the infant cry audio data to be used as input for the ViTs to investigate and evaluate the performance of the ViTs using these various features. These features include spectrogram, mel spectrogram, waveform, chroma, MFCCs, GFCCs, and tempogram, each offering unique insights into the audio signals. The spectrogram visually represents the spectrum of frequencies in a signal as it varies over time [17]. The mel spectrogram applies a mel-scale filter bank to the spectrogram, emphasizing how humans perceive sound [18]. Waveform captures the raw amplitude of the audio signal over time [15]. Chroma features capture the twelve different pitch classes, making them useful for tasks involving harmonic content [19]. MFCCs represent the short-term power spectrum of a sound [20], and GFCCs more accurately model human auditory processing [21]. The tempogram represents the variation in tempo (speed) over time [22]. GFCC and mel spectrogram features were chosen for their effectiveness in capturing audio characteristics relevant to pathology classification. GFCCs mimic the human auditory system's frequency sensitivity, while mel spectrogram enhances perceptual relevance, making both features suitable for detecting subtle variances in infant cries associated with specific health conditions [18,21].

Each of the audio feature visualizations from Figure 3 was used individually with the ViT model to explore its impact on classification performance. Each feature contributes unique information about the audio signal. GFCCs were selected for their ability to closely model human auditory perception by focusing on intensity in the frequency bands most relevant to distinguishing subtle differences in infant cries. They are highly effective in medical and speech-related tasks due to its biological alignment with human hearing, though they can be computationally more intensive. Tempogram, which captures rhythmic or tempo variations, is valuable for analyzing cyclic patterns or temporal dynamics in the cries. However, it may not be as informative for tasks where rhythmic patterns are less critical.

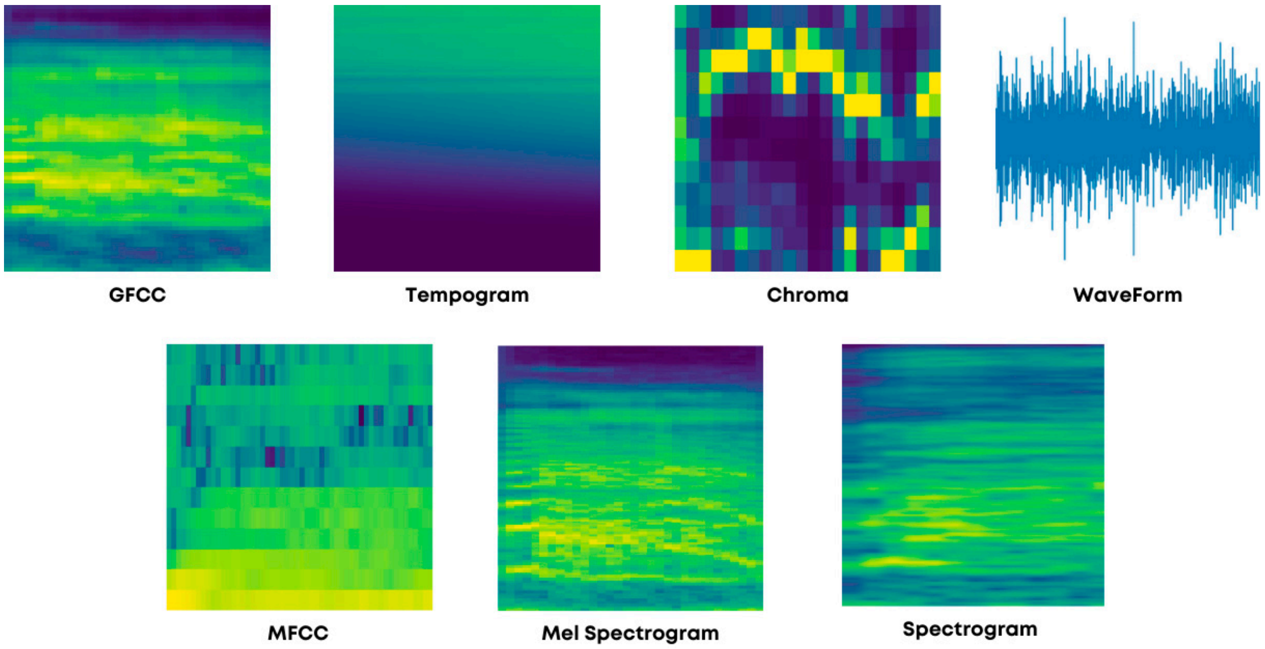

Figure 3. Visualization of extracted audio features.

Chroma features were chosen for their ability to capture pitch content, including harmonic and tonal elements, which can provide insights into the tonal characteristics of infant cries. While effective for harmonic analysis, chroma features may be less useful when pitch is not a primary aspect of the signal. Waveform captures the raw amplitude of the audio signal over time, preserving fine temporal details, making it suitable for tasks requiring high temporal resolution. However, raw waveforms can introduce noise and be more difficult for the model to process compared to spectral features. MFCC, a standard in speech recognition, was selected for its compact representation of the short-term power spectrum, emphasizing the phonetic components of the signal. Its strength lies in speech-related tasks, although it may struggle with capturing long-term dependencies.

The mel-spectrogram was included for its ability to apply a mel-scale filter to the spectrogram, aligning with how humans perceive sound via emphasizing energy distribution across the frequencies most relevant to human hearing. However, this can sometimes result in a loss of detail in higher frequency ranges. Finally, the spectrogram provides a comprehensive time–frequency representation, capturing both spectral features and temporal changes in the signal. While it is powerful, the sheer amount of data it presents can be computationally demanding.

By using each feature individually, this study evaluates how well the ViT model can leverage different representations of the audio signal for classification tasks. Previous studies have demonstrated the effectiveness of these features, such as GFCCs and spectrograms used in infant cry diagnosis systems [14], MFCCs for infant cry recognition [23], and chroma for music genre classification [24]. Tempograms have been applied to music structure segmentation [25], mel-spectrogram to audio forgery detection [26], and waveform to deep learning of music features [27]. These references underscore the validity and effectiveness of using these features independently in machine learning models for audio classification tasks.

3.3. Data Preprocessing

All extracted feature images were resized to 224 × 224 pixels to remain consistent and compatible with the structure of the ViT model. Resizing ensured that the input dimensions were compatible with the model's requirements, making the processing and analysis operations seamless.

The preprocessing involved a few important steps to extract the audio feature images and prepare them for input to the ViT model. The data preprocessing stage consists of resizing, normalization, and data splitting. Normalization was applied to ensure that all input data are on a similar scale, which helps the model converge faster and perform better. The resized images are normalized with each pixel value, along with the mean and standard deviation of the ImageNet dataset (as the ViT model was pre-trained on ImageNet). The normalization equation applied is as follows:

$$NormalizedValue = \frac{x-\mu}{\sigma}, \tag{1}$$

where x represents the original value, μ denotes the mean, and σ is the standard deviation applied to standardize data for analysis.

The dataset was split into three subsets, i.e., training, validation, and testing, to evaluate the performance and generalization of the model. The model was trained on a dataset that comprised 80% of the data, with the remaining 20% used for hyperparameter tuning to prevent overfitting and for the final purpose of testing the model when it is ready for production. These preprocessing steps ensured that the data were standardized and properly prepared for the ViT model to train and evaluate accurately and efficiently.

3.4. Vision Transformer

Given this critical role of the ViT model, in combination with previous researchers that have employed the ViT with audio data [28,29], we used a ViT model that attempts to revolutionize the field of computer vision by incorporating principles from transformers, a type of architecture that was developed for natural language processing and has since become the new standard for a variety of language-processing tasks. It introduces an original way to move from old CNNs to new state-of-the-art transformer-based models with excellent results. The key idea of ViT is to process images in patches to preserve fine-grained and long-range structures within visual data [1].

This allows transformers, known for their sequential processing capabilities, to deal with image data outside of the context of dimensionality while simultaneously bypassing the need for predefined spatial feature extraction, an essential factor when working on more intricate visual abstractions. Moreover, their unique design allows for parallel processing, which speeds up training time drastically compared to CNNs. The self-attention mechanism can handle connections at both local and global ranges of the input sequence, which is more effective for some real-world tasks.

3.4.1. ViTs Key Features and Architecture

The ViT model is just a regular 12-layer transformer encoder with multi-head self-attention mechanisms and feed-forward neural networks. This arrangement helps the model understand the image's global context and long-range dependencies.

The sequence of embedded patches has a unique class token prepended. This token gathers data from all patches and is fed into the output classification module. Patch embeddings include position embeddings (necessary to preserve spatial information in vision tasks).

As shown in Figure 4, the input image is divided into fixed-length patches (typically 16 × 16 pixels for the timm/vit_base_patch16_224 implementation) using the ViT model. The patches are then flattened and associated with a vector, which is used to feed the transformer as input tokens. The final output of the transformer encoder is fed through a classification head to make predictions. This design takes advantage of the transformer's strengths in image classification, making it easy to implement and ready for transfer learning, as shown in Figure 4 and Table 2.

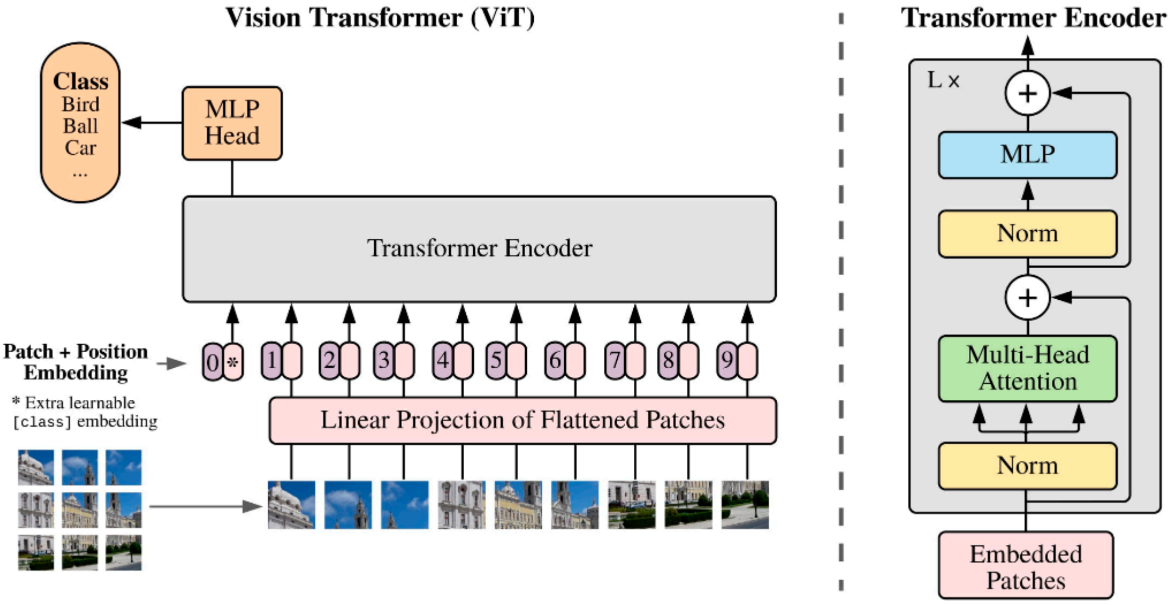

Figure 4. Vision transformer (ViT) architecture [1].

Table 2. Model specifications.

Feature	Specification
Model Name	ViT-Base Patch16 224
Patch Size	16 × 16 pixels
Image Size	224 × 224 pixels
Number of Layers	12 transformer encoder layers
Hidden Size	768
Number of Attention Heads	12
MLP Size	3072
Total Parameters	Approximately 86 million

The timm/vit_base_patch16_224 model, a ViT available in the timm library, designed for processing 224 × 224 images by dividing them into 16 × 16 patches, has shown exceptional performance across various image recognition tasks, often outperforming traditional CNNs when trained on large-scale datasets. Its applications include image classification, object detection, and semantic segmentation. ViTs are highly scalable, utilizing large datasets and significant computational resources to improve performance. The self-attention mechanism effectively captures the global context compared to local convolutions, and ViTs can be adapted to various vision tasks with minimal modifications.

In contrast to traditional audio features, which capture raw spectral characteristics like frequency, amplitude, and time-based patterns, vision transformers (ViTs) construct a 'semantic space' by learning high-level representations of the data. This semantic space enables the model to abstract and group sounds based on contextual similarities, rather than just physical attributes. While audio features depict the raw energy distribution and frequency content of a signal, the ViT model transforms these into abstract tokens that represent meaningful patterns, enabling more robust classification.

The ViT model is used in this study as it is an expert pattern recognition model for multiclass classification of infant cry data. Extensive variability in audio feature images like spectrograms, mel spectrograms, GFCCs, MFCCs, waveform, tempogram, and chroma features helps improve the accuracy of classification of pathology classes like healthy, sepsis, and RDS by the ViT model. In addition, XAI methods such as LIME, LRP, and attention mechanisms are also considered to interpret why the model classified the results as such, improving the transparency and interpretability of the classification results. This work not only leverages the remarkable properties of ViTs for handling medical audio images but also investigates their ability to highly enrich the processing of medical audio datasets into more advanced forms of analysis and classification in healthcare studies.

3.4.2. Model Training and Evaluation

This section outlines the training procedure and evaluation metrics used for the ViT model to classify infant cry audio feature images. The model was trained using the 'timm.create_model' function with the vit_base_patch16_224 architecture, pretrained on ImageNet, and adapted to our specific number of pathology classes that are listed in our dataset.

The training process was carried out using a set of predefined hyperparameters with numerous tests to identify the optimal hyperparameters. The model was initialized with pretrained weights and adjusted for the required number of output classes, as shown in Table 3. A learning rate of 7.9×10^{-5} was chosen due to its balance between speed and accuracy to prevent a global minimum. The weight decay, which acts as the L2 norm, was set to 6×10^{-3} to avoid overfitting and ensure the generalization of new data, while choosing a batch size of 32 helped capture sufficient data in each training step. The model was fully converged within 60 epochs, as shown later in the validation and training loss graphs.

Table 3. Hyperparameters of the transformer model.

Hyperparameter	Value
Number of Epochs	60
Learning Rate	7.9×10^{-5}
Batch Size	32
Weight Decay	6×10^{-3}
Optimizer	AdamW
Scheduler	OneCycleLR

The CrossEntropyLoss function was employed for its effectiveness in multi-class classification, and the AdamW optimizer was selected for its effective handling of weight decay. To optimize memory usage and computational efficiency, mixed precision training was applied using PyTorch's GradScaler. Additionally, a OneCycleLR scheduler was used to dynamically adjust the learning rate throughout the training process, facilitating better convergence and preventing the model from getting trapped in local minima.

$$CrossEntropy = -\sum_{i=1}^{M} y_i \cdot log p_i, \tag{2}$$

where M is number of classes, y_i is the true label, and p_i is the predicted probability.

The training loop, which iterated over the specified number of epochs, followed a structured process. The model was set to training mode during each epoch, and training data were loaded in batches. The optimizer gradients were zeroed, and forward and backward passes were performed with mixed precision. The loss was scaled, the gradients were computed, and the optimizer step was updated. After each batch, the learning rate scheduler was stepped. The training loss was accumulated and logged for each epoch, ensuring a systematic tracking of the training progress.

Hyperparameters were fine-tuned by starting with default parameters based on prior ViT studies. We gradually adjusted values like the learning rate, batch size, and weight decay, observing how the model converged with different numbers of epochs, optimizers, and schedulers. We performed multiple experiments to determine the optimal settings based on the model's convergence behavior and performance metrics.

Cross-validation was applied to ensure robust performance. The dataset was divided into k-fold subsets, with the model trained on k-1 folds and validated on the remaining fold. This approach ensured that the model was not overfitting to any particular subset and provided a better evaluation of its generalization capabilities.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

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

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{6}$$

where *TP* is the true positives, *TN* is the true negatives, *FP* is the false positives, and *FN* is the false negatives.

These metrics were calculated at the end of each epoch for the validation dataset. We observed the accuracy, recall, precision, and F1-score for various audio features across the validation split during the training process. The best-performing model, based on the highest validation accuracy, was saved for further training. These evaluation metrics allowed us to closely monitor the model's performance and ensure that the results were robust and reliable.

The ViT model achieved high accuracy and moderate robustness across a range of classification metrics of infant cry audio feature images. The model was able to adequately discern between healthy, sepsis, and RDS categories. The evaluation process proved that this model can be implemented in this complex classification problem.

3.5. Explainable AI

This section delves into the XAI techniques applied to interpret the ViT model's decisions in classifying infant cry audio feature images. The goal is to enhance the transparency and comprehensibility of the model's predictions, which is crucial for clinical applications.

3.5.1. Local Interpretable Model-Agnostic Explanations (LIME)

LIME was utilized as one of the primary XAI techniques. LIME is designed to explain individual predictions by locally approximating the complex model with an interpretable one. The method, used in a previous study [10], involves perturbing the input data and observing the resulting changes in the model's predictions, thereby identifying the features that most significantly influence the outcome. By applying LIME, we could determine the contribution of various audio features (such as spectrograms, mel spectrograms, GFCCs, MFCCs, waveforms, tempogram, and chroma features) to the classification results.

The optimization objective of LIME can be formally represented as

$$\xi(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{7}$$

where $\xi(x)$ is the explanation, for instance, x; f is the original black-box model; and g is the interpretable model (e.g., a linear model) from the family of interpretable models G. The term $L(f, g, \pi_x)$ is the loss function that measures how well g approximates f in the locality defined by π_x, which is the locality around the instance x, typically defined by a kernel function. Finally, $\Omega(g)$ is a complexity term to ensure that the interpretable model g remains simple preventing overfitting or overly complex explanations. The kernel function for computing the weights is

$$\pi_x(z) = exp\left(-\frac{distance(x, z)^2}{\sigma^2}\right), \tag{8}$$

where σ is a kernel width parameter that controls the locality. This function ensures that points closer to the instance x have a more significant influence on the explanation, making the model's interpretation more relevant and focused.

3.5.2. Transformer Interpretability Beyond Attention Visualization

In addition to LIME, we employed a novel interpretability method tailored for transformer models [30]. This approach addresses the limitations of existing attention-based visualization techniques by offering a more robust relevance propagation framework. The process assigns local relevance scores based on the Deep Taylor Decomposition principle. It propagates these scores through the transformer layers, considering the complexities introduced by self-attention mechanisms and skip connections.

This method integrates relevance and gradient information to produce class-specific visualizations, overcoming non-linearity challenges and maintaining overall relevance across layers [30]. This technique enables a detailed examination of which parts of the input image (converted from audio features) contribute most to the model's predictions. Combining this approach with LIME ensured comprehensive and accurate explanations for the ViT model's decisions.

LRP, a popular XAI technique, was used to unpack the decisions of our model. LRP propagates the prediction score backward through the network layers to assign a relevance score to each input feature. This method effectively decomposes the prediction, returning the output to the input features in their contributions to the final decision. Using LRP allowed us to identify which parts of the input (e.g., regions in spectrograms or types of audio features like MFCC or chroma) the model relied on to make its classifications. These relevance scores allowed us to look deeper into the model's decision-making and begin to understand its behavior [30].

The basic LRP rule for a fully connected layer can be expressed as

$$R_j^{(l)} = \sum_k \frac{a_j^{(l)} w_{jk}^{(l,l+1)}}{\sum_{j'} a_{j'}^{(l)} w_{j'k}^{(l,l+1)} + \epsilon \cdot sign\left(\sum_{j'} a_{j'}^{(l)} w_{j'k}^{(l,l+1)}\right)} R_k^{(l+1)}, \tag{9}$$

where $R_j^{(l)}$ is the relevance score of neuron j in layer l, $R_k^{(l+1)}$ is the relevance score of neuron k in layer $l+1$, $a_j^{(l)}$ is the activation of neuron j in layer l, and $w_{jk}^{(l,l+1)}$ is the weight connecting neuron j in layer l to neuron k in layer $l+1$. The term ε is a small stabilizer to avoid numerical issues. LRP, along with LIME and attention-based methods, ensures the ViT model's decision-making process is transparent and interpretable, enhancing reliability and trustworthiness in classifying infant cries.

This equation ensures that the contribution of each neuron in a given layer to the final decision is systematically traced back through the network. By assigning relevance scores, LRP makes it possible to understand which features the model considers most important in its decision-making process. The stabilizer ϵ helps prevent division by tiny numbers, which could otherwise change the relevance scores.

3.5.3. Attention Mechanisms

Attention XAI was used with a ViT in a previous study to classify chest X-ray images [10]. We integrated attention mechanisms as one of the critical XAI techniques to further improve the interpretability of the model. The mechanism helps the model attach different attention levels to different parts of the input data, which tends to represent the part that the model focuses on to make predictions. Similarly, attention maps allowed us to visualize which parts of an audio signal spectrogram (or visual representation of a waveform) of the signal carried the most weight in a model's decisions. These attention maps tell us which audio features (e.g., pitch modulations, frequency components, temporal patterns, etc.) the model focuses on when determining the sound type, which can be easily interpreted for the classification results.

3.5.4. Visualization of Model Decisions

The interpretability methods were applied to visualize the model's decision-making process in classifying infant cries into pathology classes, including healthy, sepsis, and RDS. The visualizations generated by LIME, LRP, and the transformer-specific relevance propagation method provided insights into how the model distinguishes between different conditions based on the audio features.

For example, using LIME, we identified specific segments of the spectrogram or mel spectrogram that significantly impacted the model's classification. This allowed us to understand which features are the most important in predictions. Similarly, the relevance propagation method revealed the inner workings of the attention mechanisms, highlighting the patches of the input image that the ViT model focused on when making predictions.

By employing these advanced XAI techniques, we ensured that the ViT model's decision-making process was transparent and interpretable, enhancing the reliability and trustworthiness of the classification results. These visual explanations not only validate the model's performance but also provide valuable insights for further improving the model and understanding the underlying patterns in infant cry audio features.

4. Experimental Results

4.1. Evaluation of ViT Model Performance Using Various Audio Features

4.1.1. Training and Validation Loss Analysis

The training and validation loss curves for the top three models provide additional insights into the model performance and generalization capabilities. For the GFCC-based model, the training and validation loss curves show a steady decrease in training loss, reaching near zero, and a corresponding decrease in validation loss, stabilizing over time. This indicates that the model effectively learns from the training data and generalizes well to unseen data, minimizing overfitting. The spectrogram-based model also shows a decreasing trend in training loss, with the validation loss stabilizing after initial fluctuations. Although there are some oscillations in the validation loss, it eventually levels off, suggesting that the model can generalize effectively despite the initial variability.

Similarly, the mel spectrogram-based model exhibits a continuous decrease in training loss and a stabilization of validation loss. The validation loss curve shows some fluctuations, but overall, it trends toward stabilization, indicating good generalization, as shown in Figure 5.

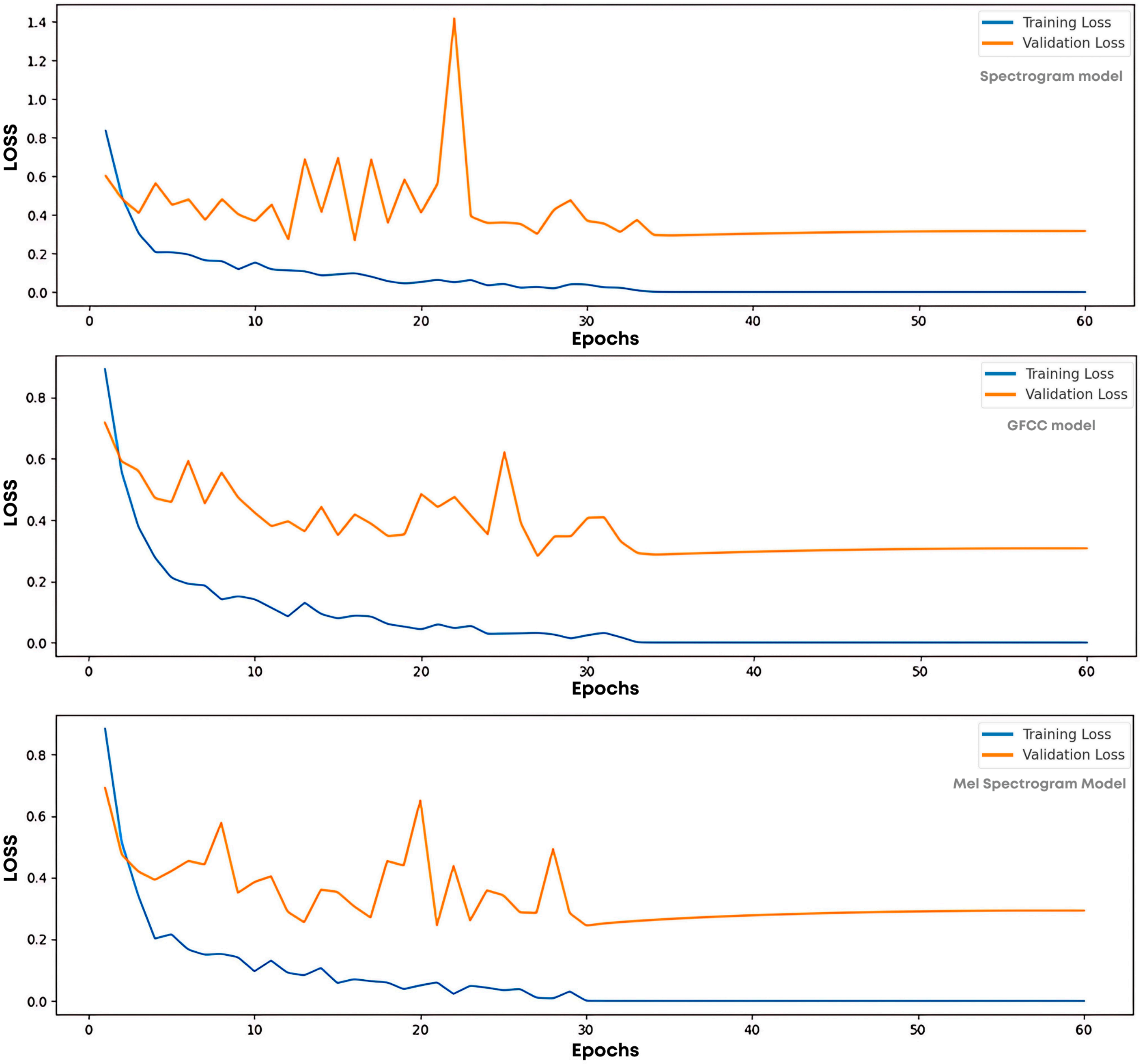

Figure 5. Training and validation loss for spectrogram, GFCC, and mel spectrogram models.

The training and validation loss curves for all three models demonstrate effective learning and generalization, with the GFCC-based model showing the most stable performance. These results, combined with the confusion matrices and ROC curve analysis, confirm that GFCC, mel spectrogram, and spectrogram features are highly effective in predicting infant pathologies, providing robust and reliable classification performance. These findings emphasize the importance of selecting and analyzing appropriate audio features for improving the accuracy and robustness of machine learning models in medical diagnosis tasks. By leveraging these insights, future research can continue to enhance the performance and applicability of such models in clinical settings.

4.1.2. Performance Metric Comparison

The experimental evaluation used a ViT model on various audio feature representations of infant crying signals. The features considered include GFCC, mel spectrogram, spectrogram, MFCC, chroma, waveform, and tempogram images. The performance of each model was evaluated in terms of accuracy, F1 score, precision, and recall. The results are summarized in Table 4.

Table 4. Performance metrics of vision transformer models on different audio feature representations of infant crying signals.

Audio Feature	Accuracy	F1 Score	Precision	Recall
GFCCs	96.33%	0.96	0.96	0.96
Mel Spectrogram	94.83%	0.95	0.95	0.95
Spectrogram	93.17%	0.93	0.93	0.93
MFCCs	80.83%	0.81	0.81	0.81
Chroma	67.83%	0.68	0.68	0.68
Waveform	63.5%	0.64	0.64	0.64
Tempogram	42.5%	0.42	0.42	0.42

Table 4 highlights that the top-performing models utilized GFCC, mel spectrogram, and spectrogram images and achieved the highest test accuracy and F1 scores. Specifically, GFCC images demonstrated the best performance with a test accuracy of 96.33% and an F1 score of 96%. This indicates a high model precision and recall level, suggesting that GFCC features are highly effective for this classification task. Similarly, mel spectrogram images also performed exceptionally well, with a test accuracy of 94.83% and an F1 score of 95%, confirming their suitability for predicting infant pathologies. Spectrogram images achieved a respectable performance with a test accuracy of 93.17% and an F1 score of 93%, highlighting their effectiveness in the classification process. Overall, GFCC, mel spectrogram, and spectrogram images achieved high F1 scores and accuracy, validating their suitability for classification. These features enabled the ViT model to leverage its ability to handle complex audio data, making these features the most accurate in classifying infant pathologies.

In contrast, other features such as MFCC, chroma, waveform, and tempogram images exhibited significantly lower performance metrics, indicating their limited utility in this specific task. MFCC images, for example, showed a substantial drop in performance with a testing accuracy of 80.83% and an F1 score of 0.81. Chroma and waveform images followed with even lower accuracies of 67.83% and 63.50%, respectively. Tempogram images, with the lowest performance metrics, demonstrated a testing accuracy of 42.50% and an F1 score of 0.42, highlighting their ineffectiveness in this context.

4.1.3. Confusion Matrix Analysis

To investigate the performance of the best three models mentioned earlier—GFCCs, mel spectrogram, and spectrogram—confusion matrices revealed the difference in correctly predicted classes on the scale (healthy, RDS, sepsis) in Figure 6. The GFCC-based model contained the lowest number of misclassification errors and the highest number of correct ones over all classes overall, with 194 healthy, 187 RDS, and 190 sepsis cases, and there was only slight confusion due to misclassification between the healthy and sepsis classes, indicating that the model performed well in distinguishing between the different pathologies. The mel spectrogram-based model performed well and predicted 194 healthy, 188 RDS, and 187 sepsis cases, which is slightly more than the GFCC model, and still, the number of misclassifications was kept low, which means it was able to predict this task. In the third order of effectiveness, the spectrogram-based model performed the highest number of misclassifications, but it is still reported as good and reliable in this task because it showed the best performance compared to the others. It also identified 187 healthy, 182 RDS, and 190 sepsis cases, as shown in Figure 6.

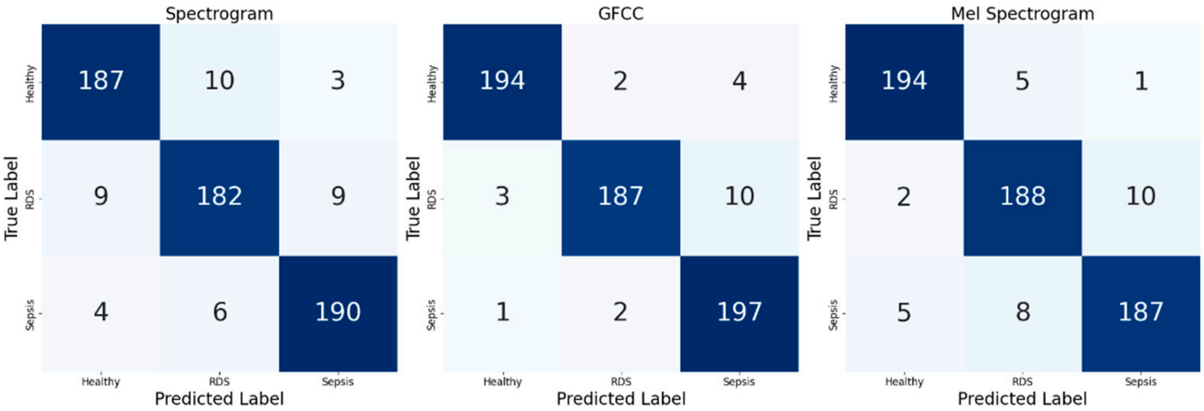

Figure 6. Confusion matrices for GFCC, mel spectrogram, and spectrogram models, respectively.

The confusion matrices underscore the superior performance of the GFCC, mel spectrogram, and spectrogram features in the context of infant pathology prediction. The high accuracy and low misclassification rates highlight the robustness of these audio features in accurately identifying different pathological conditions. These results emphasize the importance of selecting appropriate audio features for improving model accuracy and robustness in medical diagnosis tasks.

4.1.4. ROC Curve Analysis

ROC curves were generated for the top three models to further evaluate the model performance, providing a visual representation of the true positive rate versus the false positive rate for each class. The area under the curve (AUC) was calculated for each class to quantify the ability of the model to discriminate between the different conditions. As shown in Figure 6, the ROC curve for the spectrogram-based model shows AUC values of 0.95 for healthy, 0.94 for RDS, and 0.96 for sepsis, indicating a solid ability to differentiate between the classes, with particularly good performance in identifying sepsis cases. The GFCC-based model exhibited even higher AUC values, with 0.98 for healthy, 0.96 for RDS, and 0.97 for sepsis, demonstrating exceptional performance and reliability in classifying different pathologies, especially in distinguishing between healthy and sepsis cases. Similarly, the ROC curve for the mel spectrogram-based model shows AUC values of 0.98 for healthy, 0.95 for RDS, and 0.95 for sepsis, indicating high effectiveness in differentiating between the classes with consistent performance across all categories, as shown in Figure 7.

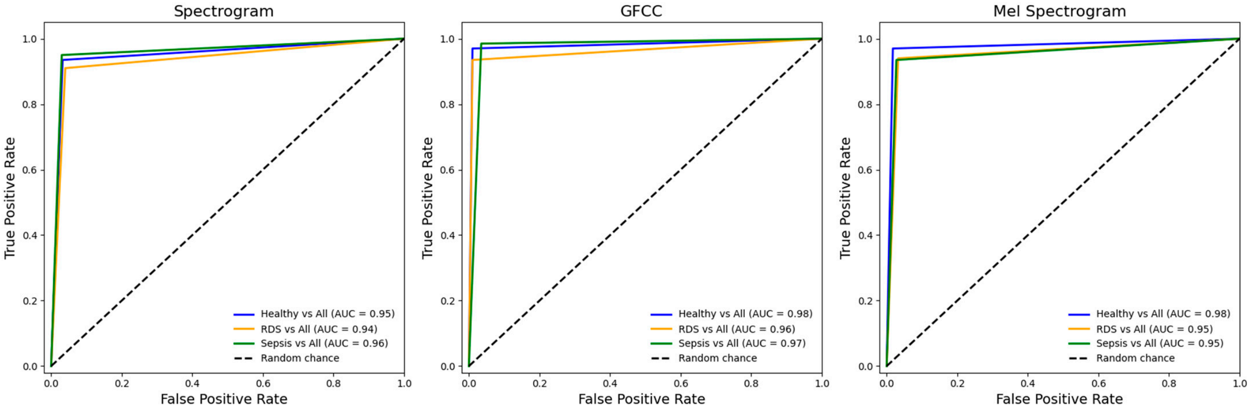

Figure 7. ROC curves for GFCC, mel spectrogram, and spectrogram models.

The ROC curves and AUC values further validate the superior performance of the GFCC, mel spectrogram, and spectrogram features for predicting infant pathologies. The high AUC values across all classes highlight the robustness and reliability of these audio features in accurately identifying different pathological conditions. These results emphasize the importance of careful feature selection in developing effective machine learning models for medical diagnosis tasks.

4.1.5. Explainable AI Results

The ViT model was analyzed using the XAI methods of LRP, LIME, and attention. It was also applied to three different audio features, i.e., GFCCs, spectrogram, and mel-spectrogram, with various accuracies. The insights are mainly based on how well the XAI methods visualized the important regions for the healthy, RDS, and sepsis classes and how these visualizations relate to the performance of the model.

Figures 8–11 illustrate how the XAI methods (LRP, LIME, and attention) highlight the critical regions within the audio features that the model relies on for classification.

LRP (Layer-wise Relevance Propagation): Brighter colors in LRP visualizations indicate areas that contribute most significantly to the model's decision. These regions often appear as concentrated vertical bands in GFCCs or spectrograms, representing specific frequency bands crucial for distinguishing between conditions.

LIME (Local Interpretable Model-agnostic Explanations): LIME provides broader, dispersed color regions, signifying approximate areas that affect the model's predictions. Unlike LRP, which pinpoints exact frequencies, LIME offers a general overview of important regions. LIME was chosen for its ability to provide interpretable, localized explanations, which are essential in medical diagnostics for building clinician trust and ensuring model transparency. By showing how changes in the input influence the model's output, LIME allows users to validate the reliability of the model's predictions.

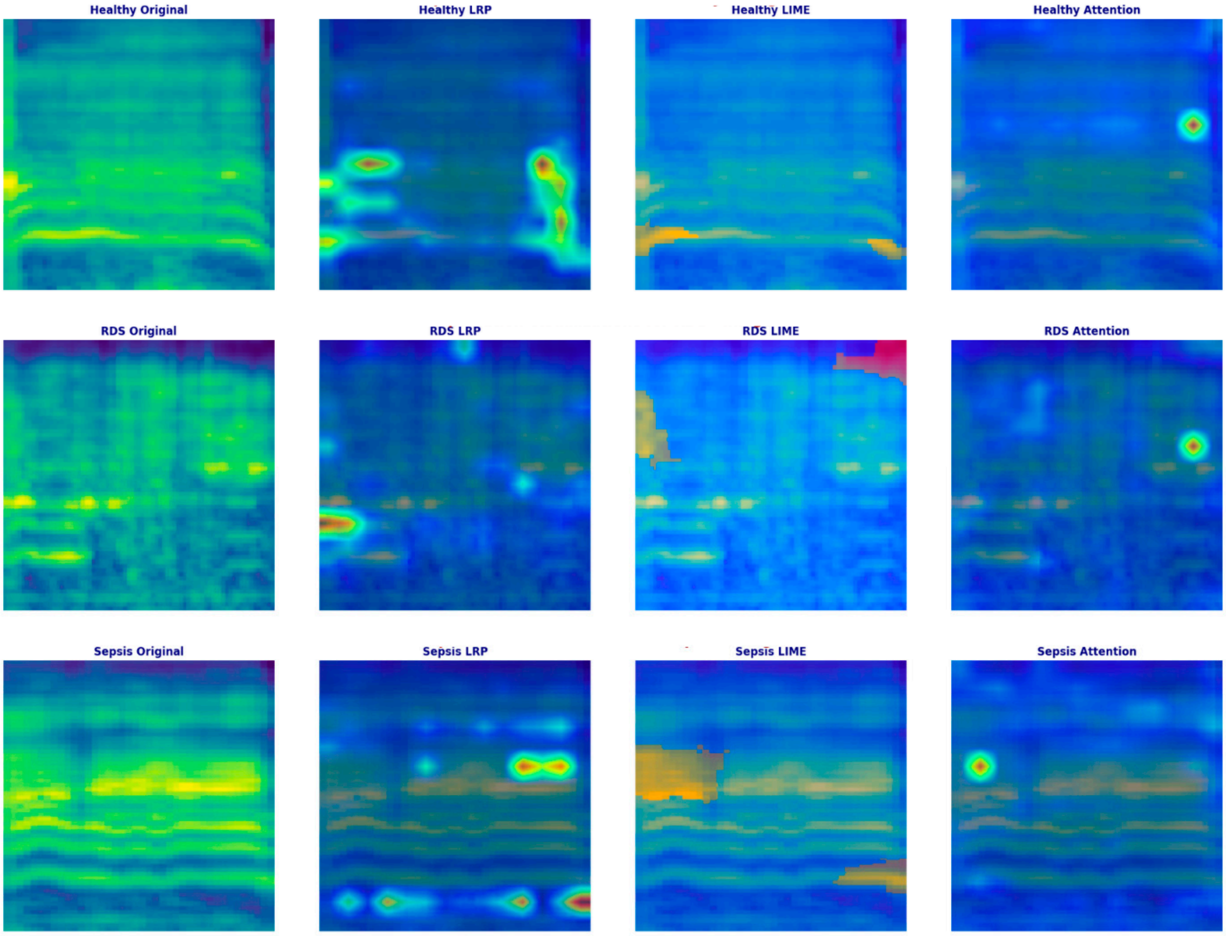

Figure 8. XAI for GFCC audio features using LRP, LIME, and attention methods.

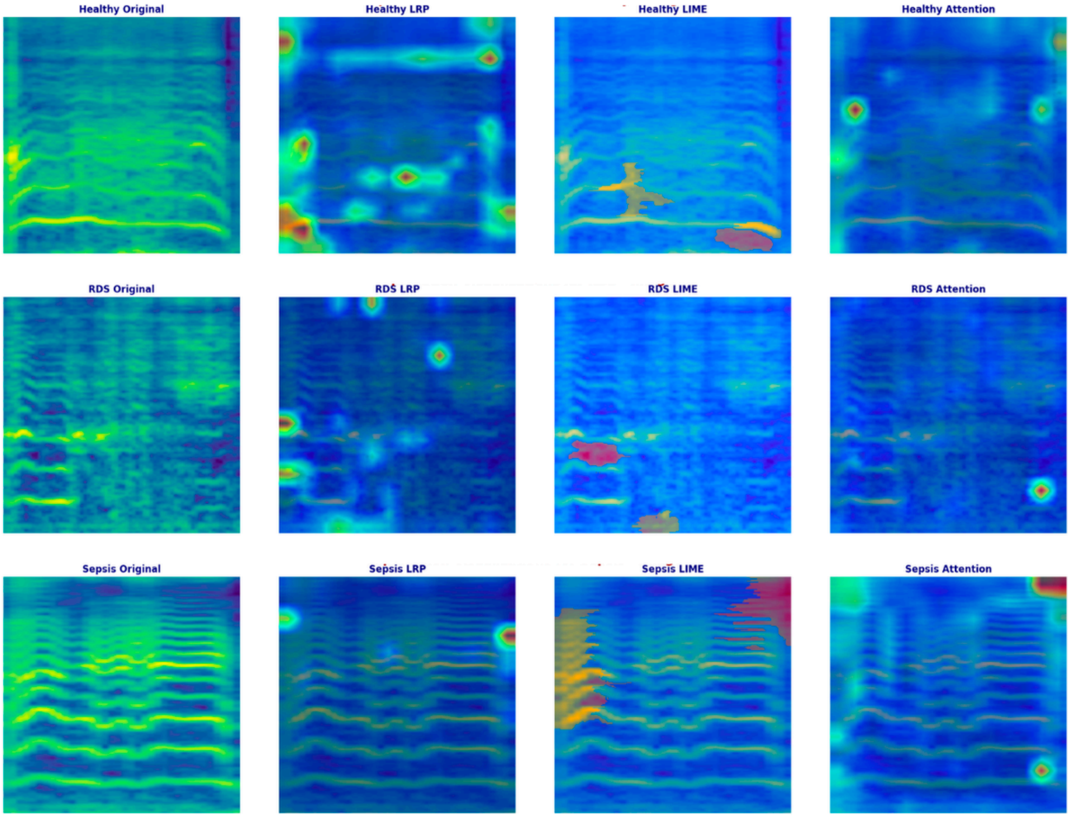

Figure 9. XAI for spectrogram audio features using LRP, LIME, and attention methods.

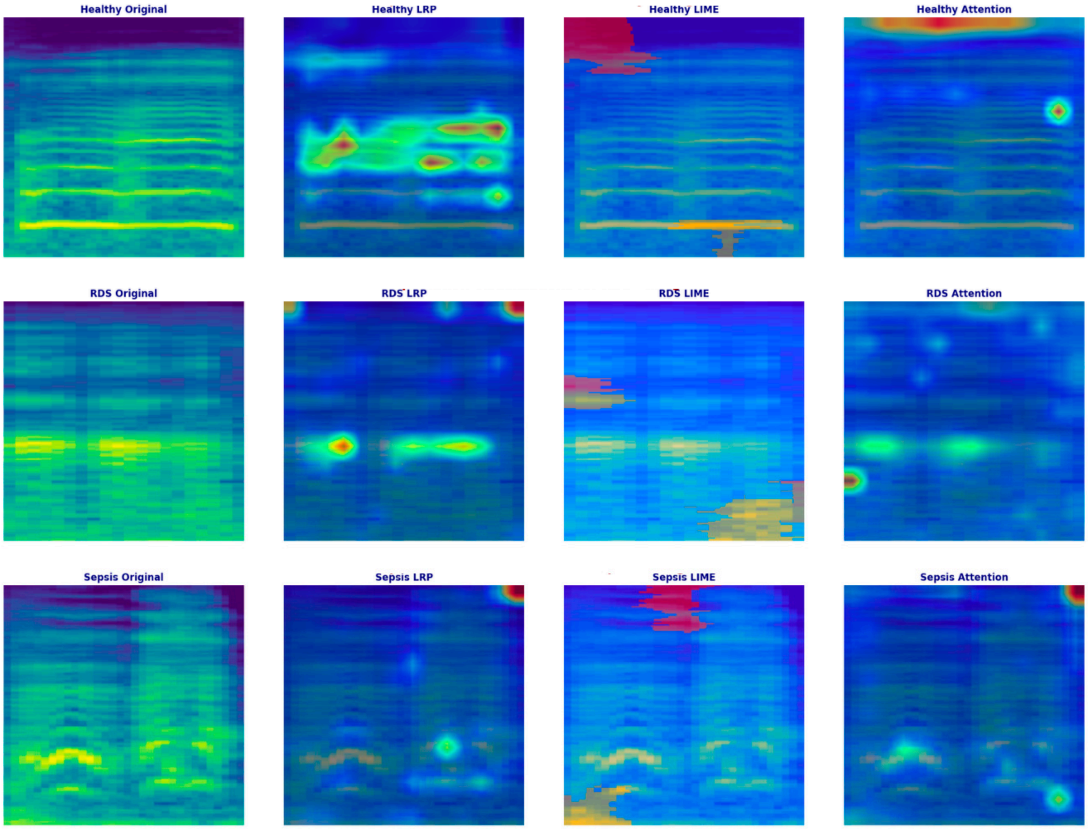

Figure 10. XAI for mel-spectrogram features using LRP, LIME, and attention methods.

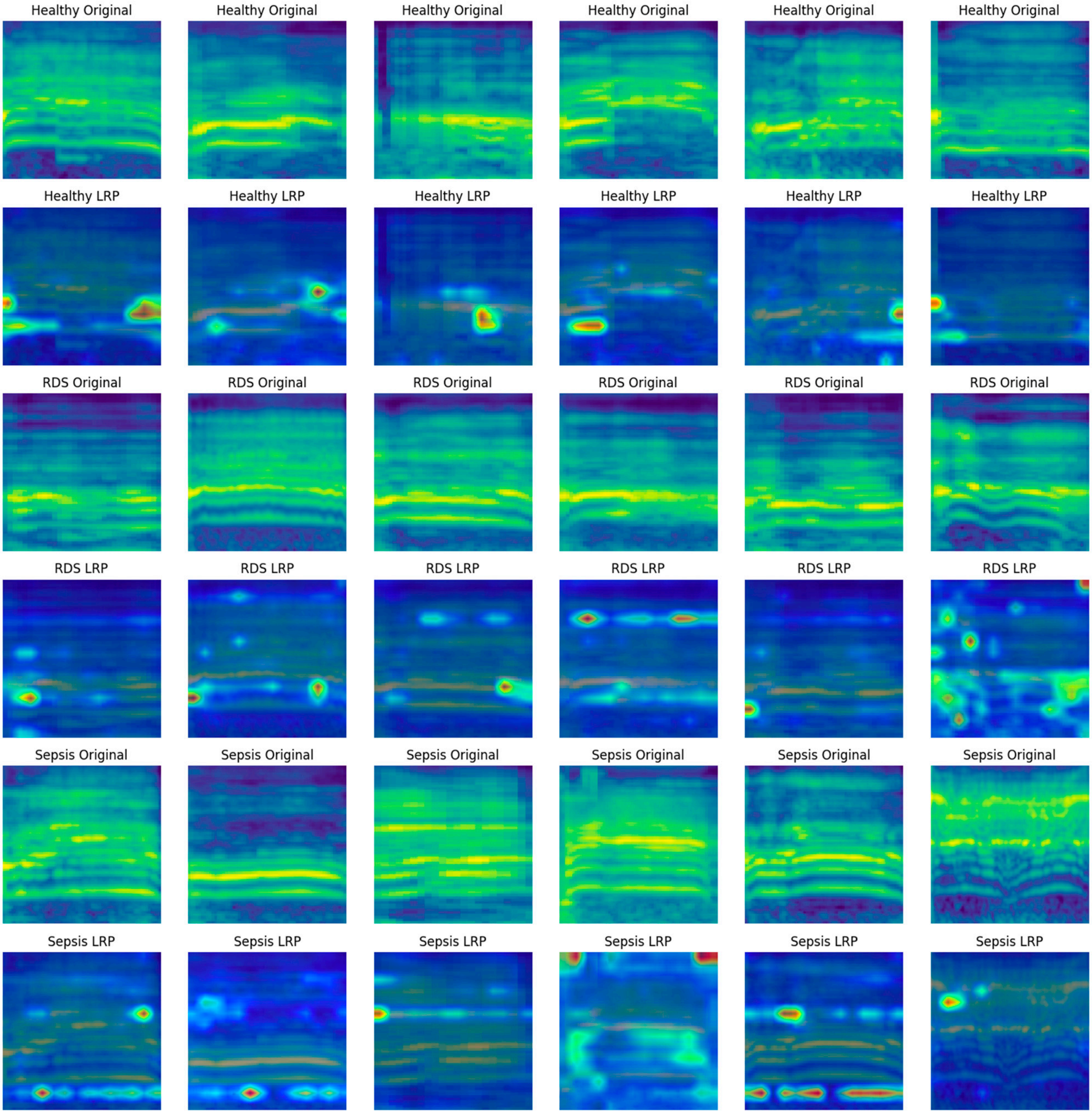

Figure 11. Visualization of model interpretations of the GFCC-ViT model using LRP.

Attention maps: Attention-based visualization highlights focal points within the audio features which the model ‘attends’ to the most during classification. This can be observed as specific, sharp points in the maps, indicating the most relevant parts of the audio spectrogram that influence the model’s decision for each class.

In the case of the GFCC feature, which achieves a remarkable 96.33% accuracy, LRP consistently reflects some vertical bands for all classes. Brighter colors in the LRP visualizations represent areas of higher importance, typically appearing as concentrated vertical bands in specific frequency ranges critical for each class. The areas of interest for the healthy class are localized in the different frequency bands, which reflect their importance. The RDS pathology also shows a vertical band structure similar to the synthetic class, which arises when specific frequencies are more important for the model’s decisions. Within the sepsis class, the bands are less uniformly dense and cover more specific frequency ranges, corresponding to the representation the model has learned to prioritize within these ranges. In contrast, LIME outputs broader, dispersed color regions, indicating general areas that influence the model's predictions rather than pinpointing specific frequencies. This provides a general idea of how important each feature is in making the decision for part classification. The attention maps for GFCCs highlight sharp focal points, showing where the model concentrates most during classification. These focal points are more distinct in GFCCs, reflecting a high precision of interpretation, as shown in Figure 8.

Figure 9 shows the LRP visualizations for spectrogram features, which obtained an accuracy of 93.17%, showing vertical bands similar to GFCCs and specifically indicating the relevance of frequency. Brighter colors signify higher relevance areas in the frequency bands. The healthy class would instead have some clean vertical bands, indicating high regions of interest that stand out as the most important frequencies. RDS shows more diffuse highlights than the healthy class but is also more specific, implying valuable frequencies. Sepsis class bands are more ambiguous because they overlap, producing less distinct, sparser bands that would visually indicate a wider, less fine-tuned range of frequencies. LIME's broad regions highlight approximate areas contributing to model decisions across all classes, providing a global approximation of attribute importance. Attention maps for spectrogram features depict clear focal solid points in the healthy class, sharp but scattered highlights in the RDS class, and precise but more dispersed foci for the sepsis class, showing a generalized focus.

Figure 10 shows the LRP visualizations for mel-spectrogram features, which obtained an accuracy of 94.83%, showing prominent vertical bands similar to GFCCs, marking frequency ranges that are most significant for model decisions. Brighter colors indicate regions of higher relevance. The healthy class has very clean vertical bands, which stand out as key frequencies. RDS shows more diffuse highlights, implying valuable frequencies, while the sepsis class has overlapping, sparser bands, suggesting a broader range of frequencies the model focuses on. LIME shows broader, distributed areas across all classes, indicating approximate importance over larger regions. Attention maps for mel-spectrogram reveal specific focal points for the healthy class, sharp but scattered highlights for RDS, and more dispersed foci for sepsis, indicating the model's generalized focus on relevant frequency bands.

LRP methods provide the most focused and fine-grained visualizations for all individual features. It searches for specific frequency bands that can be vital to the model's decisions. In contrast, LIME offers a more extensive perspective on the importance of different features over large regions of the audio signals. There was a good balance of attention, as seen in the attention maps: clear focal points with some spreading out represent where the model thinks about most.

Overall, LRP provides fine-grained details, with bright spots signifying critical frequency bands; LIME offers a broader perspective on feature importance, and attention maps highlight specific focal points, enhancing interpretability across the different audio features. This comprehensive approach using multiple XAI methods ensures the transparency and reliability of the model's classification decisions.

Figure 11 shows how explainable AI classifies the three pathology classes via LRP on GFCC audio features. The six figures are pairs of original GFCC spectrograms and LRP maps, indicating parts of the input that have the most critical impact on the model predictions. The GFCC features were chosen over other features because of their highest accuracy with the ViT model. This visualization helps us understand how the model separates various classes and highlights its decisions' key informative areas. The effectiveness of these XAI methods depends critically on the accuracy measurement of each feature type when using the ViT model. Models with higher accuracy standards, like GFCCs, provide much shorter and sharper explanations. Conversely, low-accuracy models based on other audio features cause non-uniform visualization. This indicates that the performance of the XAI method is influenced by the model's performance, and more transparent explanations correspond to a better-performing model.

Across these pathologies, the clarity and focus vary for each XAI method and feature type. This shows important details of how the model understands different audio features differently. These findings explain the strengths and weaknesses of each XAI method according to their consistency in model performance, which is particularly inspiring in the cases of complex models such as ViT, as shown in Figure 11.

The regions of interest depend on the feature types and the methods related to XAI, and these change in their clarity and focus for each pathology, indicating the subtlety in the way audio features are interpreted by our model. These analyses show the importance and shortcomings of each XAI approach to each other and the actual model, which gives concrete explanations of interpretability aspects of deep models like ViT.

5. Discussion

In this research, we aimed to study the effects of image-based audio embeddings on classification tasks for ViTs and XAI. The experimental results demonstrate the effectiveness of ViTs in various audio features (GFCCs, mel-spectrogram, and spectrogram), as we can observe high classification accuracy using these methods. These feature representations were then compared with each other, and the GFCC-based model performed the best with an accuracy rate of 96.33%, followed by mel-spectrogram (94.83%) and spectrogram (93.17%). This shows how the chosen audio feature representation significantly affects its performance for classification tasks with ViTs. GFCCs perform better overall because they capture the most important aspects of the audio signal that are needed to separate between different pathologies in newborns and how human auditory processing occurs. GFCCs' ability to capture the phonetic nuances of infant cries, such as subtle frequency variations related to different pathologies, enhances their effectiveness in distinguishing between healthy and pathological cries. In contrast, the mel-spectrogram and spectrogram provide broader frequency ranges, but they may not isolate the critical frequency bands as effectively, leading to less precision in differentiating pathologies. The high classification performance of the mel spectrograms could also be partially attributed to the fact that they also encode properties related to how humans perceive sound frequencies. MFCCs, chroma, waveform, and tempogram did not perform as well, suggesting that they are less useful for this specific task.

The results obtained from earlier study models [3,14,16,31,32] designed to classify various audio data of infant cries into different pathologies highlight the significance of the proposed model in this research, as shown in Table 5.

Table 5. Comparison of performance metrics across different studies and the proposed model.

Metric	Model [16]	Model [14]	Model [31]	Model [32]	Model [3]	Proposed Model
Classes	2 classes	3 classes	3 classes	4 classes	3 classes	3 classes
Audio features	GFCCs, HR	GFCCs, HR, spectrogram	Spectrogram	Linear Frequency Cepstral Coefficients (LFCCs)	Spectrogram	GFCCs
ML algorithm	Multilayer perceptron	Fusion deep learning (CNN)	SVM + CNN	XGBoost	Transformer	Transformer
Accuracy	95.92%	97.50%	92.50%	92%	98.69%	96.33%
Precision	95%	97.51%	88.80%	-	98.73%	0.96
Recall	95%	97.53%	89.30%	-	98.71%	0.96
F1 score	95%	97.52%	88.90%	92.30%	98.71%	0.96

Each model utilizes different approaches and combinations of audio features and ML algorithms for classifying infant cries into different pathologies. Model [16] uses a multilayer perceptron (MLP) algorithm along with audio features such as GFCCs and heart rate (HR), achieving a high accuracy of 95.92% with a simplified classification of two classes. Model [14] employs fusion DL using CNN, GFCC, HR, and spectrogram features, achieving an impressive accuracy of 97.50% for classifying into three pathologies. Model [31] combines SVM with CNN for classification, using spectrogram features and achieving an accuracy of 92.5% for three classes. Model [32], which utilizes XGBoost with Linear Frequency Cepstral Coefficient (LFCC) features, reaches an accuracy of 92% for four classes. Model [3] employs the spectrogram feature, which uses an audio spectrogram transformer (AST) and achieves an impressive accuracy of 98.69% for three classes. Finally, in this study, the proposed model utilizes the GFCC feature, employs a ViT, and achieves an accuracy of 96.69% for three classes, demonstrating the significance of using the ViT model. From the perspective of previous studies, however, the advancement of the proposed model and model [3] lies in their capability to achieve high results, suggesting that transformers could be the best model in audio classification tasks. In comparing models for classifying infant cries into various pathologies, the transformer model proposed in this study and in [3] outperform previous research regarding accuracy, F1, precision, recall, and score.

One of the critical aspects of our experiment is the ease of use and excellent performance of ViT compared to more complex models used in existing works. For instance, [14,16] are rich, complex approaches, including a deep learning model-based flat first impression monitor that achieves extremely high accuracy with heavier settings. In contrast, our experiments demonstrate that using the ViT architecture combined with only one feature type per model run (GFCC, mel-spectrogram, spectrogram), our proposed simple method achieves competitive or superior performance. This simplifies the feature extraction task and makes the model more aromatic. In the case of our study, feature extraction relies on transforming audio signals into images (e.g., spectrograms, mel-spectrograms, GFCC) before being processed by the ViT. The vision transformer (ViT) demonstrated strong performance compared to the audio spectrogram transformer (AST), likely due to its ability to capture global dependencies in image-based audio features through self-attention. This architecture enables effective processing of high-dimensional data with reduced complexity. Although AST processes raw audio data and achieves slightly higher accuracy, it requires significantly more computational resources and training time. Thus, ViT provides a balanced approach with competitive accuracy and simplified feature extraction.

XAI methods such as LRP, LIME, and attention made ViT predictions in this study much more interpretable. LRP identifies which parts of the audio are crucial for the model's predictions, helping users understand why the model makes certain decisions by producing pixel-level visualizations that highlight the exact frequency bands where deeper models tunnel their predictions. LIME shows how changes in the input affect the model's output, helping users assess the model's reliability across different scenarios. It provides a comprehensive view of feature importance as the input data are perturbed and the model's predictions are analyzed before and after those changes. Attention maps reveal which parts of the audio the model focuses on, ensuring the model is concentrating on the most relevant features. These insights make the model easier to interpret and use effectively, providing a better understanding of the key audio components driving the classifications.

This has wide-ranging implications for medical diagnostics, where near-perfect accuracy and interpretability in the ViT models of image-based audio representations such as stethoscope recordings were noted. Realizing this potential, we are now focused on developing a classification system of infant cry sounds that can be used as an early diagnostic tool for effective neonatal healthcare. XAI techniques provide transparency that enables healthcare professionals to trust and understand the model's predictions, thus promoting the uptake of these AI-based tools in clinical care.

This study's limitations are as follows: Although the results of this study are very promising, it has some limitations. One limitation is that the embryo dataset may need to fully capture the range of natural variability in infant cries of different populations and conditions. In the future, more diverse and representative samples should be explored to supplement the dataset. In addition, investigation into other advanced audio feature extraction techniques and hybrid models may be beneficial to enhance classification performance. Optimizing the ViT architecture and training process to work better with more extensive and more complex datasets is also future work. An insight into the joint analysis of multimodal data, i.e., from applications such as audio and vision fusion, could be a more holistic study for medical diagnostics. Further development of XAI methods is important to ensure that explanations generated by these models are transparent, consistent with the underlying model logic structure, and valuable for clinical practice.

Overall, this study shows the effectiveness of incorporating vision transformers for image-based audio representation to classify infant cries and achieve high accuracy with features such as GFCC and mel-spectrogram. Incorporating such explainable AI techniques helped us gain important insights into the model's decision-making process. Also, it improved transparency and trust, with a significant increase in verifiability and replicability. These results imply that data-driven models like AI-based tools have the potential to assist in medical diagnostics, providing us with significant leverage to facilitate widespread health screening; however, the same result also draws attention to the necessity for selecting rather than creating features and understanding those we want if robust and trustworthy models are required.

6. Conclusions and Future Work

This research highlights the performance of the ViT model in classifying infant crying signals by converting audio signals into image-based representations. The GFCC feature provided the best results among various features, achieving an accuracy of 96.33% and an F1 score of 96%. This highlights the critical role of audio feature representation in enhancing the performance of ViTs, particularly in classifying infant pathologies. Integrating XAI techniques, such as LRP, LIME, and an attention layer, improved the transparency and interpretability of the ViT model's predictions. Among these, LRP appeared as the most effective XAI method for determining the three pathologies (healthy, sepsis, and RDS), providing the models' most detailed and accurate visualizations. This level of interpretability is crucial for medical applications, where understanding the model predictions builds trust and ensures reliability. Overall, the ViT approach showed significant advantages over other complex models, simplifying the processing pipeline while delivering accurate classification results. The combination of high performance and enhanced interpretability makes this approach a promising tool for medical diagnostics, offering both precision and transparency in critical healthcare applications.

Several areas could be improved for ViT models to demonstrate better performance and applicability in medical diagnostics after future studies have been conducted. The data need to be updated for more excellent coverage and more generalizability. Hopefully, this approach will capture a larger cross-section of infant crying sounds from many cities, countries, and conditions, thus improving the model's generalizability. Moreover, using other sophisticated feature extraction from audio profiles and ensemble models can further enhance the classification results. This can be improved either by combining them with several other features or by deriving new feature extraction mechanisms specifically for the given pathologies. For example, modeling multimodal data integration, like combining audio and visual signals, could provide a more holistic way of medical diagnosis, incorporating the surrounding elements that single modal models might fail to consider.

Finally, optimizing the ViT architecture and training pipelines to deal with larger datasets should also be considered in the future. Playing around with different numbers of layers, attention heads, or embedding dimensions can also help to refine the model's performance and efficiency. We need continuous refinement of XAI techniques to ensure that model explanations remain interpretable, precise, and actionable in clinical decision-making. The main limitation we have with this is that, by definition, the interpretability of audio-controlled input–output pairs cannot be stripped down from many types of XAI methods. Hence, there will always be an opportunity for some XAI method to develop new or better explanation methods that can help us understand the how and why at least a little bit more. Lastly, the models need to be implemented in practice on real-life patients. This also includes the front end of user interfaces, data privacy and security, and clinical trials to test the effectiveness of the model in real-world applications.

Author Contributions: Conceptualization, S.M. and M.T.; methodology, S.M. and M.T.; software, S.M. and M.T.; validation, A.H. and C.T.; formal analysis, A.H.; investigation, S.M.; resources, A.H. and C.T.; data curation, S.M.; writing—original draft preparation, S.M. and M.T.; writing—review and editing, A.H. and C.T.; visualization, S.M.; supervision, A.H. and C.T.; project administration, A.H. and C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee, École de Technologie Supérieure #H20100401. Approval date: 4 March 2022.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not publicly available due to the restrictions imposed by the ethics committee of Ecole de Technologie Superieure.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., Eds.; Curran Associates, Inc.: Nice, France, 2021; pp. 12116–12128. Available online: https://proceedings.neurips.cc/paper_files/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html (accessed on 20 November 2024).
2. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
3. Tami, M.; Masri, S.; Hasasneh, A.; Tadj, C. Transformer-Based Approach to Pathology Diagnosis Using Audio Spectrogram. *Information* **2024**, *15*, 253. [CrossRef]
4. Khalil, M.; Khalil, A.; Ngom, A. A comprehensive study of vision transformers in image classification tasks. *arXiv* **2023**, arXiv:2312.01232.
5. Böhle, M.; Fritz, M.; Schiele, B. Holistically explainable vision transformers. *arXiv* **2023**, arXiv:2301.08669.
6. Verma, P.; Berger, J. Audio transformers: Transformer architectures for large scale audio understanding. Adieu convolutions. *arXiv* **2021**, arXiv:2105.00335.
7. Bazi, Y.; Bashmal, L.; Al Rahhal, M.M.; Al Dayil, R.; Al Ajlan, N. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
8. Chetoui, M.; Akhloufi, M.A. Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-rays. *J. Clin. Med.* **2022**, *11*, 3013. [CrossRef] [PubMed]
9. Hossain, S.; Chakrabarty, A.; Gadekallu, T.R.; Alazab, M.; Piran, J. Vision Transformers, Ensemble Model, Transfer Learning Leveraging Explainable AI for Brain Tumor Detection and Classification. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 1261–1272. [CrossRef] [PubMed]
10. Komorowski, P.; Baniecki, H.; Biecek, P. Towards Evaluating Explanations of Vision Transformers for Medical Imaging. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023. [CrossRef]
11. Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.-R. Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer Nature: Berlin/Heidelberg, Germany, 2019; pp. 193–209. [CrossRef]
12. Park, N.; Kim, S. How do vision transformers work? *arXiv* **2022**, arXiv:2202.06709.
13. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
14. Zayed, Y.; Hasasneh, A.; Tadj, C. Infant Cry Signal Diagnostic System Using Deep Learning and Fused Features. *Diagnostics* **2023**, *13*, 2107. [CrossRef] [PubMed]
15. Dieleman, S.; Schrauwen, B. End-to-end learning for music audio. In Proceedings of the ICASSP 2014—2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 6964–6968. [CrossRef]
16. Khalilzad, Z.; Hasasneh, A.; Tadj, C. Newborn Cry-Based Diagnostic System to Distinguish between Sepsis and Respiratory Distress Syndrome Using Combined Acoustic Features. *Diagnostics* **2022**, *12*, 2802. [CrossRef] [PubMed]
17. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]
18. Choi, K.; Fazekas, G.; Sandler, M.; Cho, K. Convolutional recurrent neural networks for music classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2392–2396. [CrossRef]

19. Ellis, D.P.; Poliner, G.E. Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, Signal Processing, Honolulu, HI, USA, 16–20 April 2007; pp. IV-1429–IV-1432. [CrossRef]
20. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
21. Ayoub, B.; Jamal, K.; Arsalane, Z. Gammatone frequency cepstral coefficients for speaker identification over VoIP networks. In Proceedings of the 2016 International Conference on Information Technology for Organizations Development (IT4OD), Fez, Morocco, 30 March–1 April 2016; pp. 1–5. [CrossRef]
22. Grosche, P.; Muller, M.; Kurth, F. Cyclic tempogram—A mid-level tempo representation for musicsignals. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010, Dallas, TX, USA, 14–19 March 2010; pp. 5522–5525. [CrossRef]
23. Liang, Y.-C.; Wijaya, I.; Yang, M.-T.; Juarez, J.R.C.; Chang, H.-T. Deep Learning for Infant Cry Recognition. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6311. [CrossRef] [PubMed]
24. Shi, L.; Li, C.; Tian, L. Music Genre Classification Based on Chroma Features and Deep Learning. In Proceedings of the 2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP), Marrakesh, Morocco, 14–19 December 2019; pp. 81–86. [CrossRef]
25. Tian, M.; Fazekas, G.; Black, D.A.A.; Sandler, M. On the use of the tempogram to describe audio content and its application to Music structural segmentation. In Proceedings of the ICASSP 2015—2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 419–423. [CrossRef]
26. Dincer, S.; Ustubioglu, B.; Ulutas, G.; Tahaoglu, G.; Ustubioglu, A. Robust Audio Forgery Detection Method Based on Capsule Network. In Proceedings of the 2023 International Conference on Electrical and Information Technology (IEIT), Malang, Indonesia, 14–15 September 2023; pp. 243–247. [CrossRef]
27. Natsiou, A.; O'Leary, S. Audio representations for deep learning in sound synthesis: A review. In Proceedings of the 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA), Tangier, Morocco, 30 November–3 December 2021; pp. 1–8. [CrossRef]
28. Lu, L.; Liu, C.; Li, J.; Gong, Y. Exploring Transformers for Large-Scale Speech Recognition. *arXiv* **2020**, arXiv:2005.09684.
29. Li, Y.; Tagliasacchi, M.; Rybakov, O.; Ungureanu, V.; Roblek, D. Real-Time Speech Frequency Bandwidth Extension. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 691–695. [CrossRef]
30. Chefer, H.; Gur, S.; Wolf, L. Transformer Interpretability Beyond Attention Visualization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
31. Vincent, P.M.; Srinivasan, K.; Chang, C.-Y. Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. *Front. Public Health* **2021**, *9*, 670352. [CrossRef]
32. Joshi, V.R.; Srinivasan, K.; Vincent, P.M.D.R.; Rajinikanth, V.; Chang, C.-Y. A Multistage Heterogeneous Stacking Ensemble Model for Augmented Infant Cry Classification. *Front. Public Health* **2022**, *10*, 819865. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.