ML Audio Classification of Parent-Infant Interactions

Gwen Powers, Elisabeth Waldron, Will Sivolella

***Abstract*— In this study, we investigate the application of machine learning (ML) techniques to classify audio recordings of parent-infant interactions within the Latinx communities. Our goal is to identify distinctive vocal patterns to inform research on early Autism interventions. Employing a dataset comprising over 200 hours of recorded audio from 32 different Latinx families, we analyze these interactions using two advanced ML models: convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). The models were trained to differentiate between vocalizations from parents, infants, and background noises. Our results demonstrate that our models achieve classification accuracies of 67% and 69% for CNN and LSTM respectively, with the LSTM model showing particular promise in capturing temporal vocal patterns. These findings underscore the potential of machine learning in early diagnostic settings, offering a scalable tool for healthcare professionals to enhance early intervention strategies. The study also discusses the implications of these technologies in real-world applications, emphasizing the importance of robust data preprocessing and model tuning to handle the nuanced variations in naturalistic audio data.**

I. INTRODUCTION

A. Purpose and Background

The purpose of this study is to apply advanced machine learning techniques to classify and analyze parent-infant vocal interactions in Latinx households. By distinguishing between distinct vocalizations of parents, infants, and background noises, this research aims to inform early autism research and interventions within the Latinx community. This project supports the broader mission of the University of Virginia's Supporting Transformative Autism Research (STAR) initiative, aiming to improve the quality and accessibility of care for Latinx families with members on the autism spectrum.

The UVA STAR initiative operates within the School of Education and Human Development and is dedicated to interdisciplinary research aimed at improving outcomes for children with autism. A significant barrier in autism intervention, particularly in Latinx communities, is the late diagnosis due to cultural nuances, language barriers, and a general lack of specialized services.

Our project draws inspiration from the LENA (Language ENvironment Analysis) technology, which uses sophisticated audio analysis to monitor and evaluate the natural conversational turns between children and adults. Although effective, the high cost and complexity of LENA limit its accessibility for many families and community programs. Furthermore, our focus on the Latinx community addresses the urgent need for culturally sensitive diagnostic tools that consider linguistic diversity and cultural specificity, which are potentially overlooked in existing models.

The importance of this study lies in its potential to bridge gaps in early detection and intervention within underserved communities, ultimately contributing to more equitable health outcomes and enriching the field of developmental disorder diagnostics through technological innovation.

B. Data Description

The dataset employed in this study consists of approximately 200 hours of MP3 audio recordings from 32 different Latinx families residing in Central Virginia. These recordings capture the natural daily interactions between infants, aged between 6 months and 2 years, and their parents or caretakers. The audio was collected as part of a community-based participatory research initiative, aiming to inclusively gather data reflective of the everyday [uncontrolled] environments and interactions these children experience. Each family contributed at least one hour of audio, recorded during various daily activities such as meals, playtime, and other family interactions, ensuring the capture of lax behavior and communication.

An important component of our dataset includes 10 hours of metadata. This data consists of manually classified timestamps by student researchers to help define distinct vocalizations. These time stamped labels serve as the ground

truth for training our supervised machine learning models.

This study included a multi-class target variable with the primary goal of classifying a parent or infant speaking. The predictors include a range of acoustic features such as the frequency, amplitude, and power of the audio signals extracted with Short-Time Fourier Transform (STFT). We took an additional approach that uses a sophisticated feature extraction, mel-frequency cepstral coefficients (MFCCs - became optional). These acoustic features are extracted to provide insights into the pitch, tone, and rhythm of the vocalizations, which are indicative of the interactive patterns and potential developmental concerns.

By focusing on a specific demographic group, we aim to develop a model that is finely tuned to the cultural and linguistic nuances present in the vocal interactions of Latinx families, thereby improving the specificity and utility of our diagnostic tool.

II. METHODOLOGY

A. Data Preprocessing

The initial step in our data preprocessing involved converting the raw MP3 audio files into two formats: CSV for detailed numerical analysis and image formats for convolutional neural network processing. This dual-format approach allows us to leverage different types of analysis—spatial and sequential—enhancing the robustness of our findings. Following the mp3 conversion, the audio files were segmented according to metadata tags, which include timestamps and speaker identification. This segmentation was critical for aligning the data with corresponding labels accurately, facilitating precise model training and validation.

Key acoustic features were extracted from each audio segment using two primary techniques: Short-Time Fourier Transform (STFT). STFT was employed to analyze the frequency and time properties of the audio signals, providing a spectrum that shows how these properties vary over time. Feature extraction is essential for our models as they encapsulate the fundamental attributes of the sounds within the recordings.

To ensure that the inputs into our machine learning models were consistent and free from scale discrepancies, we implemented a standardization procedure. All numerical features extracted were normalized to have zero mean and unit variance. This normalization process is particularly important when dealing with features like frequency and amplitude which can vary significantly across recordings, depending on the recording environment and the speaker's distance from the microphone.

In addition to data value standardization, padding was deployed to the data matrices to standardize their dimensions that contained varying segment shapes. This step was essential for batch processing in neural networks, ensuring that each input had the same shape. For image-formatted data, this involved adjusting the image dimensions to a fixed size, whereas for the csv data, we added zeros to shorter sequences after defining the maximum segment length. Proper alignment and padding are crucial for maintaining the integrity of temporal relationships within the data during model training.

B. CNN Model Development and Training

For our study, we chose the ResNet50 architecture as the base for our convolutional neural network due to its proven efficacy in image and pattern recognition tasks, particularly in complex datasets. The Resnet50 model, renowned for its deep architecture and residual learning framework, facilitates the training of deeper networks by mitigating the vanishing gradient effect. This model was initially pretrained on a large dataset (ImageNet) to capture a wide range of features, which provides a strong foundational knowledge for our task.

The top layers of the model were customized to suit our specific classification needs. This customization involved adding several new layers: a global average pooling layer to reduce the dimensionality of the feature maps and avoid overfitting, a dense layer to interpret the features extracted by the convolutional layers and learn non-linear combinations of these features, and a softmax output layer that outputs the probabilities of the three classes (parent, infant, background noise), providing a clear classification.

The training of our CNN model was carried out in two phases. During the initial feature extraction phase, we froze the weights of the pretrained Resnet50 layers to leverage the learned features without altering them. This approach allows the model to adapt to the specifics of our audio data without losing the generic patterns learned from ImageNet. After the feature extraction, we unfroze some of the top layers of the Resnet50 model and continued training. This fine-tuning

phase was crucial for allowing the model to adjust more deeply to the particular characteristics of our dataset.

We used the Adam optimizer, known for its efficiency and adaptive learning rate capabilities, which helps in converging faster and more effectively. The loss function used was categorical cross entropy, ideal for multi-class classification problems like ours.

To ensure the model performed optimally, we conducted extensive hyperparameter tuning, experimenting with different learning rates, batch sizes, and numbers of epochs. We utilized a validation split from our dataset to monitor the model's performance during training. The model's performance was evaluated based on its accuracy and the F1-score on the validation dataset.

C. RNN Model Development and Training

The Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN), was selected for its superior ability to model time-series data and capture long-term dependencies, which are essential in audio processing. The LSTM model is particularly effective for tasks where context and the temporal sequence of data points are crucial for accurate predictions.

For our specific application—classifying segments of parent-infant vocal interactions—the LSTM architecture was designed as follows: an input layer receives sequences of feature vectors extracted from the audio files, specifically the STFTs that encapsulate temporal dynamics. Multiple layers of LSTM units were used to process the input sequences, allowing the model to learn from the temporal dependencies across the audio samples and BatchNormalization Layer were incorporated to stabilize and accelerate the network training. Dropout layers were implemented to prevent overfitting by randomly ignoring a subset of neurons during training and a dense layer was used to ensure fully connections between the inputs and outputs of the LSTM layers, facilitating classification among the three categories: adult, infant, and background noise before an output softmax activation function was used to provide the probabilities for each class, ensuring a robust classification output.

The training of the LSTM model involved several key steps. We initialized the model weights with small random values to ensure symmetry breaking in the learning process. The Adam optimizer was chosen for its adaptive learning rate capabilities, which help in efficiently converging to the optimal weights, and categorical cross entropy was used due to its effectiveness in multi-class classification tasks, measuring the disparity between the predicted probabilities and the actual class labels.

Hyperparameter tuning played a critical role in optimizing the LSTM model. Parameters such as the number of LSTM units, the learning rate, and the batch size were adjusted based on the performance observed during validation. We also employed techniques like early stopping by monitoring the validation loss over a specific number of epochs to terminate training when the model performance no longer improved, and to prevent overfitting. For the RNN, we utilized a validation split from our dataset to monitor the model's performance during training as well. The model's performance was evaluated based on its accuracy and the F1-score on the validation dataset similarly as we did with the CNN.

D. Model Enhancements

One of the first improvements involved segmenting the data into training, validation, and testing sets by decreasing the length of the segments, utilizing metadata obtained from the audio recordings. This approach ensured that the model could be trained on a diverse subset of data while being validated and tested on separate, unseen datasets. This segmentation helps in assessing the model's ability to generalize to new data, a crucial factor in real-world applications.

Extensive hyperparameter tuning was conducted for both models to optimize their performance. This process involved careful tuning of the learning rate for both models to ensure they converge to optimal solutions without overshooting or getting stuck in local minima. Next, while Adam optimizer was initially chosen for its adaptive learning rate features, we experimented with other optimizers like SGD and RMSprop to compare performance impacts. Finally, adjusting the batch size and the number of epochs during training helped in finding the right balance between computational efficiency and model accuracy.

To combat overfitting—a common problem in deep learning models—we implemented early stopping mechanisms for the LSTM model. Training would halt if the validation loss did not improve for a predefined number of

epochs. This approach saves training time and computational resources while preventing the model from learning noise in the training data. Regularization techniques such as Dropout for the LSTM and L2 regularization for the CNN were also utilized to penalize large weights and encourage the learning of simpler models that generalize better to new data.

Enhancements also included the integration of advanced statistical tools to better analyze model performance. F1-Scores and confusion matrices were particularly useful for understanding the models' strengths and weaknesses in classifying different classes. The F1-score, a harmonic mean of precision and recall, provided a better sense of model accuracy when dealing with imbalanced classes. To address any class imbalance in the training data, under-sampling of the over-represented classes and over-sampling of under-represented classes were employed. This ensured that the models did not bias towards the majority class.

III. RESULTS

A. CNN Findings

The CNN model, utilizing a Resnet50 architecture adapted for audio classification, achieved a classification accuracy of 67%. While this accuracy is promising, especially given the complexity of the audio data and the nuances of parent-infant interactions, it also suggests areas for improvement. The model's loss, measured at 0.76, indicates the average error per classification instance, pointing to the need for further model tuning to enhance prediction reliability.

The weighted average F1-score for the CNN model stood at 39%. This relatively low score compared to the accuracy suggests a discrepancy between precision and recall, indicating that while the model is relatively accurate overall, it may be failing to correctly identify all positive samples (low recall) or is misclassifying negatives as positives (low precision). This is particularly critical in the context of detecting subtle differences in vocal patterns, which are essential for early detection of developmental issues.

The confusion matrix provided deeper insights into the model's performance across the three classes: parent, infant, and background noise. The matrix highlighted that while the model performed reasonably well in distinguishing background noise, it struggled more with differentiating between parent and infant vocalizations. This challenge is likely due to the similar acoustic properties shared between parent and infant speech, especially in a naturalistic setting where overtalk and background sounds can overlap.

Graphical plots of train and validation loss versus epochs illustrated that the model's training process was stable, with a gradual decrease in training loss. However, the validation loss plateaued early, suggesting that the model might be underfitting the data, possibly due to a lack of complexity or insufficient training epochs. Similarly, train and validation accuracy graphs showed an initial rapid improvement, which then stabilized, underscoring the potential need for additional features or more complex model architectures.

TABLE I. CONVOLUTIONAL NEURAL NETWORK

Type	**RESNET50**		
	Epochs	***Accuracy***	***Loss***
Train	10	0.73	0.65
Validation	10	0.67	0.76

Figure 1. Training results of ResNet50 model

B. RNN Findings

The RNN model, LSTM, demonstrated a slightly higher overall accuracy of 69% compared to the CNN model. This improvement underscores the LSTM's capability to handle temporal and sequential data more effectively, which is crucial in analyzing the dynamic and time-dependent nature of audio signals. The model reported a loss of 0.65, indicating a better average error rate per instance than the CNN, reflecting its enhanced ability to model the complex relationships in the data.

The weighted average F1-Score for the LSTM model was 60%, a substantial improvement over the CNN model. This higher F1-score suggests a better balance between precision and recall, a critical factor when classifying audio segments where the distinction between parent and infant vocalizations can be subtle. The improved recall indicates that the LSTM model is capable of capturing more true

positive cases, essential for applications like early developmental disorder detection where missing a positive case can have significant implications, and even a misdiagnosis of autism which can lead to incorrect or unnecessary treatment.

From the confusion matrix, it was evident that the LSTM model excelled particularly in distinguishing infant vocalizations from background noise, a key requirement for analyzing infant-led interactions. However, similar to the CNN, it faced challenges with the parent class, where the model occasionally confused parental words with background noises, especially in overlapped speech scenarios. This aspect highlights the need for further refinement in feature extraction and perhaps more targeted training data that can help the model learn these distinctions better.

Training and validation plots revealed a consistent decrease in both loss metrics over epochs, with a smoother convergence compared to the CNN. This suggests that the LSTM was effectively learning from the training data without significant overfitting issues, likely due to its ability to manage sequence dependencies better. Accuracy plots also showed an upward trend, supporting the robustness of the model in handling the dataset.

Further investigation into the model's performance indicated that LSTM was particularly adept at recognizing patterns over longer sequences, such as prolonged crying or laughing sounds from infants, which are vital cues for developmental assessment. However, rapid shifts in sound types, such as quick exchanges in conversation, were sometimes missed, pointing to potential areas for improving the model's responsiveness to fast-changing audio inputs.

TABLE 2. RECURRENT NEURAL NETWORK

Type	**LSTM**		
	Epochs*	***Accuracy***	***Loss***
Train	60	0.61	0.75
Validation	60	0.69	0.65

* early stopping implemented for the RNN model, based on the lag of the validation loss over 5-10 epochs.

Figure 2. Training results of LSTM model

IV. DISCUSSION

This study aimed to leverage advanced machine learning techniques to analyze parent-infant vocal interactions, focusing particularly on the Latinx community. By employing CNN and LSTM models, our research sought to classify and interpret distinct vocal patterns which could potentially indicate developmental disorders.

The LSTM model demonstrated superior performance with a higher accuracy (69%) and F1-score (60%) compared to the CNN model. This suggests that the LSTM's ability to process sequences and maintain information across time is particularly advantageous for audio classification tasks that are inherently sequential and temporal. The LSTM model effectively captured the nuances of vocal interactions over time, which is critical for distinguishing subtle developmental cues in infant vocalizations.

The findings from this study underscore the potential of using machine learning to develop diagnostic tools that are not only effective but also accessible. Particularly in communities where linguistic and cultural nuances significantly impact the diagnosis of developmental disorders, such as in Latinx populations, the ability to analyze vocal interactions in natural settings is invaluable. These tools can help bridge the gap in early diagnosis and intervention, potentially altering developmental trajectories in a positive manner.

Despite encouraging results, both models faced challenges, particularly in distinguishing between parental speech and background noise, which could lead to misclassifications. This issue highlights the need for further refinement in feature extraction and possibly more sophisticated model architectures that can better handle overlapping sounds and the complexities of real-world audio data. However, due to the nature of the study being in an uncontrolled environment and having background noises such as television characters with similar pitched voices to the participants, the performance of the LSTM model remained promising, performing similar to related studies and the LENA software itself, both having a wide range of 50 to 80 percent accuracy, depending on the specific class being identified and task being performed.[1]

The insights gained from both models in this study, has the potentiality of guiding and advancing future research towards more specialized models that integrate both spatial

[1]Cristia A, Lavechin M, Scaff C, Soderstrom M, Rowland C, Räsänen O, Bunce J, Bergelson E. A thorough evaluation of the Language Environment Analysis (LENA) system, 2021 [5].

and temporal data processing capabilities, possibly utilizing an ensemble model approach that could pull on both methods strengths and advantages and, in theory, would hopefully reduce error, loss, and bias in contrast.

V. CONCLUSION

This study has effectively demonstrated the application of machine learning techniques, specifically convolutional neural networks (CNN) and long short-term memory networks (LSTM), in analyzing parent-infant vocal interactions within the Latinx communities. Our findings highlight the potential of these technologies to identify distinct vocal patterns that may indicate developmental disorders, offering a promising avenue for early diagnostic and intervention tools.

The LSTM model, with its superior handling of temporal sequences, achieved a classification accuracy of 69% and an F1-score of 60%, indicating its potential effectiveness in real-world applications where sequential and temporal dynamics are critical. Although the CNN model showed lower performance metrics (67% accuracy and 39% F1-score), it still contributed valuable insights into the spatial features of audio data, underscoring the complexity of audio classification tasks and the need for tailored model architectures.

The practical implications of our research are significant, especially in the context of providing scalable and cost-effective tools for early detection of developmental disorders in under-resourced settings. By leveraging machine learning, we can enhance the capabilities of existing diagnostic frameworks, potentially making early intervention more accessible to communities like the Latinx population, where such resources are often limited. These tools can help mitigate the impacts of late diagnosis and improve developmental outcomes for children at risk of developmental disorders.

Reflecting on our methodological approach, the use of both CNN and LSTM models provided a comprehensive understanding of the challenges and requirements of audio data analysis. This dual approach not only enriched our analysis but also set the groundwork for future research to explore hybrid models or other advanced machine learning techniques that could offer more nuanced insights.

In conclusion, the integration of machine learning into the field of developmental diagnostics holds considerable promise for transforming early intervention strategies. Our study serves as a foundational step towards realizing the potential of these technologies to contribute positively to public health, particularly in enhancing the lives of children and families in diverse communities.

VI. ACKNOWLEDGEMENTS

Our sincere thanks go to Dr. Michaela DuBay for her guidance and sponsorship, Dr. Aiying Zhang and Jessica Zhang for their invaluable mentorship, and the participating families for their essential contributions. We're also grateful to Professor Adam Tashman, who leads our Capstone Program, for his steadfast support and insight, and our peers and faculty at the School of Data Science at the University of Virginia that made this research possible. Our efforts reflect a collective commitment to contribute meaningfully to the field of early autism detection.

We extend our deepest gratitude to all those who contributed to the success of this study. Special thanks to Dr. Michaela DuBay for her invaluable guidance and sponsorship, and Dr. Aiying Zhang for her expert mentorship throughout the course of this research. We are also immensely grateful to Professor Adam Tashman for his support and insightful contributions, which were instrumental in the development of this project. Our appreciation extends to Jessica Zhang for her dedicated assistance and to all the participating families whose cooperation and contributions were essential to our data collection. We also thank our peers and faculty at the School of Data Science at the University of Virginia for their encouragement and constructive feedback. This project was made possible through the collective effort and commitment of everyone involved, reflecting our shared dedication to advancing the field of developmental disorder diagnostics.

REFERENCES

[1] LENA Foundation. (n.d.). Retrieved from https://www.lena.org/

[2] National Institutes of Health. (2021). PMC article on ASD. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7855224/

[3] University of Virginia Education Research Initiatives. (n.d.). Supporting Transformative Autism Research. Retrieved from https://education.virginia.edu/research-initiatives/research-centers-labs/supporting-transformative-autism-research

[4] GeeksforGeeks. (n.d.). Understanding of LSTM networks. Retrieved from https://www.geeksforgeeks.org/understanding-of-lstm-networks/

[5] Cristia A, Lavechin M, Scaff C, Soderstrom M,

Rowland C, Räsänen O, Bunce J, Bergelson E. A thorough evaluation of the Language Environment Analysis (LENA) system. Behav Res Methods. 2021 Apr;53(2):467-486. doi: 10.3758/s13428-020-01393-5. PMID: 32728916; PMCID: PMC7855224.