Using Transfer Learning, SVM, and Ensemble Classification to classify Baby Cries based on their Spectrogram Images

Lillian Le*
Institute for Artificial Intelligence
University of Georgia,
Athens, Georgia
lle@uga.edu

Abu Nadim M.H. Kabir*
Department of Computer Science
Georgia State University
Atlanta, Georgia
akabir3@student.gsu.edu

Chunyan Ji+
Department of Computer Science
Georgia State University
Atlanta, Georgia
cji2@student.gsu.edu

Sunitha Basodi+
Department of Computer Science
Georgia State University
Atlanta, Georgia
sbasodi1@student.gsu.edu

Yi Pan
Department of Computer Science
Georgia State University
Atlanta, Georgia
yipan@gsu.edu

***Abstract*—Babies cannot communicate with formal language and instead convey necessary messages through their cries. In babies, the first few months of their growth period are critical to the rest of their lives, as many conditions, such as deafness or brain damage from asphyxia, can be remedied if they are detected during this time period, preventing irreparable damage. The ability to differentiate between types of cries of a baby can prove extremely useful for parents with newborn children. To achieve this, we employ several machine learning, deep learning and ensemble classification techniques. In our work, we use transfer learning with the existing pre-trained convolutional neural network of ResNet50, a Support Vector Machine (SVM). We also perform ensemble classification to combine the predictions of the SVM and deep learning model to classify between different types of baby cries. Models are trained on spectrogram images of the audio files taken from the Baby Chillanto Database. We evaluate our models with ten iterations of 5-fold cross-validation and our models achieve accuracies of more than 90%.**

***Index Terms*—Baby cry classification, spectograms, Resnet, SVM, decision fusion**

I. Introduction

A babies' first few months are critical to the rest of their lives. Many conditions, such as deafness or brain damage from asphyxia, can be targeted and remedied if caught early on, preventing irreparable injury. However, infants cannot communicate with formal language during this crucial point in life. Instead, babies convey some reflection of physical and emotional needs through crying. To an untrained ear, a baby's cry may always sound the same, but research has shown that these cries have different meanings representing different needs of the baby. The ability to differentiate between types of cries from a baby can prove extremely useful for parents with newborn children. This development would expand and improve the process of providing responsive care and address potential medical issues.

Asphyxia and deafness are two conditions within infant children that would greatly benefit from accessible, reliable, and noninvasive detection methods. Currently, these conditions require expensive and lengthy testing to reach a conclusive diagnosis, a process which may also result in false positives or no diagnosis at all. Only 69.1% of infants who do not pass a hearing screening test are diagnosed with hearing loss before three months of age, potentially attributed to the fact that infants must undergo a variety of invasive tests to conclusively diagnose hearing loss, such as genetic testing, blood testing, and behavioral observations [1]. While 98% of American-born infants are screened for hearing loss, usually before leaving the hospital, asphyxia is not commonly tested without symptoms first being present [2]. Procedures to diagnose asphyxia can involve MRIs, CT scans, EEGs, blood tests, and metabolic exams. These diagnostic tests require time, equipment, and medical expertise. They are also not available in many areas in the world. Even if they are available, many parents may not be able to afford or access proper care for their new child. The American Academy of Pediatrics affirms that almost 36% of infants who fail their initial hearing loss screening are lost to follow-up [1]. More accessible forms of detection would prevent unneeded invasive testing, provide parents with more easily obtainable preliminary information, and hopefully reduce the rate of lost to follow-up.

Additionally, the ability to distinguish among hunger, pain, and normal crying, can be immensely useful for new and old parents alike to know the needs of their child. An accessible tool to accurately predict a child's needs from the crying signal can give parents valuable information to provide better care

*+Authors contribute equally

978-1-7281-4121-3/19/$31.00 ©2019 IEEE
DOI 10.1109/MASSW.2019.00028

of the baby. This paper focuses on classifying different baby crying signals using machine learning methods such as neural networks, SVM and Ensemble Classification.

II. Related Work

In the field of baby cry classification, there have been many attempts made in the past two decades. Many machine learning methods and feature extraction methods were used in previous works. Neural networks were used as early as 2004 when Orion F. Reyes-Galaviz & Carlos Alberto Reyes-Garcia built an infant cry recognition system, which classifies normal, deaf, and asphyxiated infants crying signals, and the classification accuracy was able to reach 86.06% [3]. SVM is also widely used in this area. Researchers have used SVM and some specialized versions to classify baby crying and the classification accuracy reached 93.8% on a specific defined database [4], [5]. M. Moharir et al. [6] classified the normal and asphyxiated classes within the Baby Chillanto Database using convolutional neural networks such as GoogLeNet and AlexNet on the waveform images converted from the audio signals. They achieved a maximum of 94% accuracy. Aomar Osmani et al. [7] implemented a variety of machine learning techniques on a dataset of 288 recordings of infant pre-cry vocalizations and cries gathered from over 30 babies of different ethnicities. The machine learning techniques implemented were an SVM, Bagged Trees, Boosted Trees, Decision Tree, kNN, and subspace kNN, where the kNN achieved the highest accuracy of 97.78%. CY. Chang et al. [8] used a dataset of 490 baby cries that they gathered from 37 infants between the ages of 1 and 10 days old to distinguish if a baby was hungry, sleepy, or feeling tired. They processed this dataset to focus more on the voices of the babies using a thresholding mechanism and then extrapolated 15 key features from the sound files. They then used a Sequential forward floating search (SFFS) algorithm and a Sequential Backward Selection algorithm to determine which features were the most important. The group ran the data through a DAG-SVM with a k-fold cross-validation method to achieve maximum accuracy of 92.17%.

III. Proposed Method

We used the Baby Chillanto Database obtained from the National Institute of Astrophysics and Optical Electronics, CONACYT, Mexico. The database contains a total of 2268 one-second samples from children ranging from newborn to nine months of age. This dataset is categorized into five cry labels as shown in Table I.

We propose a method to use transfer learning, SVM, and ensemble classification to classify the baby crying signals. The motivation is to combine deep learning models with the traditional machine learning models to improve the results.

TABLE I
Baby Cillanto Database

Infant Cry Category	Asphyxia	Deaf	Hunger	Normal	Pain
Number of Samples	340	879	350	507	192
Total	2268				

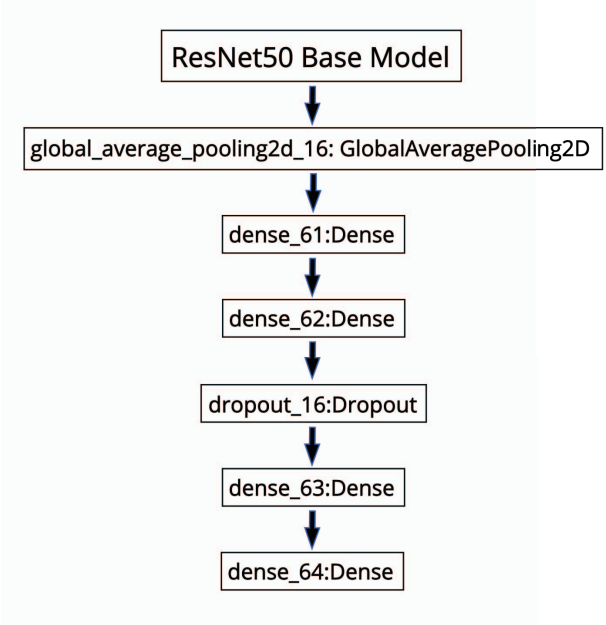


Fig. 1. TL_Resnet50 architecture

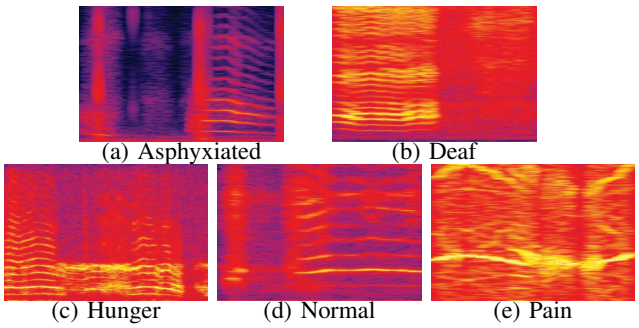
Deep learning models usually require a large amount of data and tuned architectures to determine the best configuration which can predict with higher accuracies. Since, our data is only limited to 2268 samples, we explore transfer learning techniques. In particular, ResNet models are chosen for their simplicity and better performance. In traditional machine learning models, SVM models give better performance. We experiment with both these models in different settings and combine the results using decision fusion. Four different models were implemented to compare results. We built a small convolutional neural network, a transfer learned ResNet50 based model, an SVM, and a combined model of ResNet50 and SVM for ensemble classification. The small convolutional neural network consisted of three convolutional layers with feed-forward propagation. We implemented this simple model as a baseline to compare our results against. We will refer to this model as the "Small CNN" for this paper. ResNet50 is a pre-trained 50 layer convolutional neural network trained on the ImageNet database. We selected this model due to its high accuracy, favorable speed, and ease of use, as it comes shipped with Keras, a popular neural network library. Since this model has already been trained on the ImageNet dataset and has achieved great performance when handling image classification tasks, we found it be suitable to improve the predictive capabilities of our initial "Small CNN". We used a transfer learning approach for the ResNet50 model and modified the primary ResNet50 model to predict for 5 classes rather than the original 1000 from ImageNet. Furthermore, we appended a GlobalAveragePooling layer, a dropout layer with a rate of 0.25, and four additional dense layers to the base ResNet50 model, to produce better results for our given dataset as shown in Figure 1. Within this paper, the model is referred to as "TL_ResNet50".

An SVM is a classifier that defines a plane to separate data into different categories. We elected to use an SVM for its high accuracy when handling smaller datasets. The SVM in this paper used a one-versus-rest strategy to predict for a multi-class dataset.

Once the TL_ResNet50 model and SVM were able to reach high levels of accuracy, we developed an ensemble classification method to combine their predictions. We compared the prediction matrix generated by both models for each test image to decide on a final prediction; the individual prediction values of each label in each models' prediction matrix were averaged and then stored in a temporary matrix. Finally, we recorded the maximum value in the temporary matrix for each test image, into a final prediction matrix. We explored ensemble classification to see if it could produce better results than the individual models on their own. This model is also referenced as "Combined Model" throughout the paper.

IV. Experiments and Results

From the original database of baby cries, we converted the .wav audio files into 300 dpi .png colored spectrograms of size 60x90, using the software Sox [9]. A spectrogram is a two-dimensional image in which the x-axis represents time and the y-axis represents frequency. The color shows the intensity of the energy found in a certain frequency at a certain time. The lighter the color is, the stronger the sound. Spectrograms are commonly used for feature representation of audio signals. They can show the energy level of different frequencies at all timesteps of the signals. Research in speech recognition and acoustic event detection has shown it is effective to use spectrograms for audio classification. We decided to use the spectrograms instead of the MFCC, LPCC features in our baby crying classification because we have found it to be suitable for our transfer learning model. A few sample spectrograms of different classes can be found in Figure 2. As the images show, the pain spectrogram is in yellow and red colors because babies cry much harder when they are in pain. But the asphyxiated image is a darker purple color because the sick baby's sound is weak.



(a) Asphyxiated (b) Deaf (c) Hunger (d) Normal (e) Pain

Fig. 2. Spectrogram Samples of Baby Cries.

The spectrogram images and their respective labels were first converted to numeric arrays using Keras' preprocessing library. After that, we passed the dataset through sklearn's train_test_split method, which shuffled and split the dataset into an 80:20 training and testing split with a common random seed to ensure reproducibility of the split for both models. We then separately trained and evaluated the Small CNN, TL_ResNet50, and SVM using the training dataset.

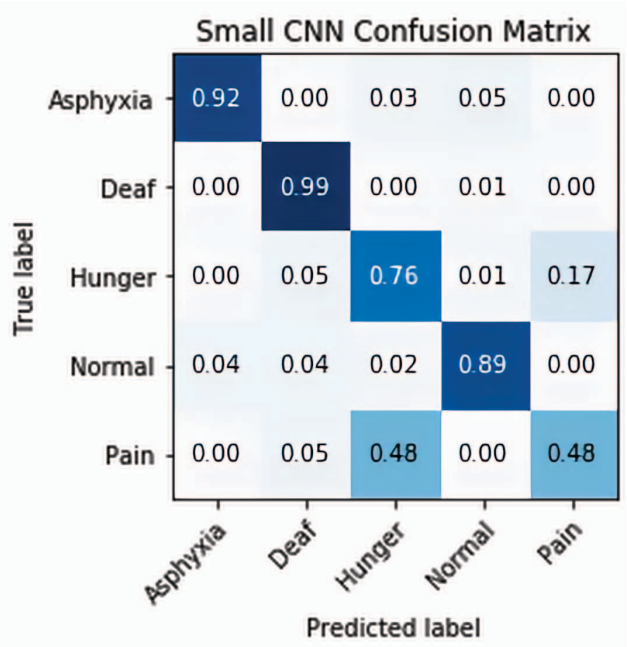


Fig. 3. Confusion Matrix of the Small Convolutional Neural Network (CNN)

During the training process, we experimented with many different learning rates, epochs, batch sizes, gamma values, kernels, and c parameters, to increase the prediction capabilities of our models. The Small CNN was trained over 10 epochs with a batch of 50 and a learning rate of 0.0001 using the Adam optimizer. The TL_ResNet50 model followed similar parameters with the exception of 3 epochs rather than 10. The SVM received its best results when gamma was set to 'scale' and the c parameter was 1000.

We initially determined the accuracy of the Small CNN and TL_ResNet50 by using the existing Keras metric, categorical_accuracy, which compares the index of the maximum argument of the true label in the labels array against the index of the maximum argument in the predictions array. SVM's initial testing accuracy is defined through the Scikit-learn function, accuracy_score, which also compares the predicted label with the true label for a given input batch. After running each model through ten iterations of 5 fold cross-validation, these are

TABLE II
Results for the Models

Model	Accuracy
Small CNN	87.03%
TL_ResNet50	90.80%
SVM	90.10%
Combined Model	91.10%

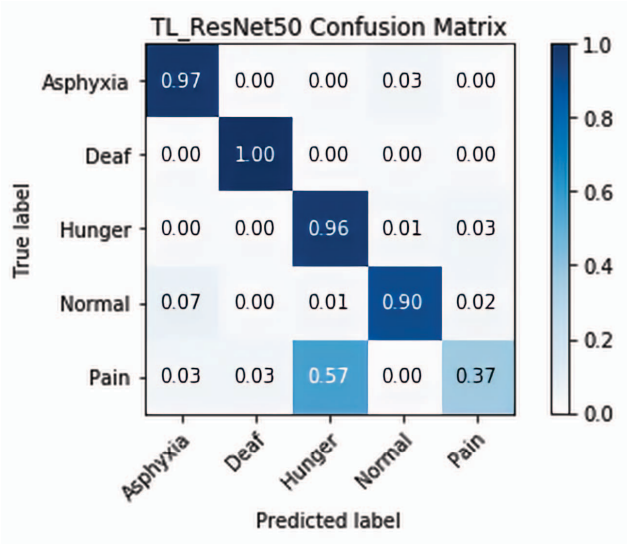


Fig. 4. Confusion matrix of TL_ResNet50 (transfer learning)

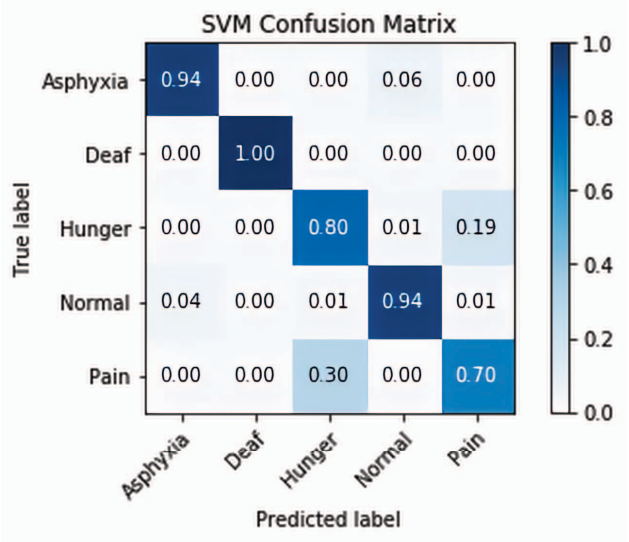


Fig. 5. Confusion Matrix of Support Vector Machine (SVM)

their respective results. The Small CNN averaged 87.03%. The TL_ResNet50 model achieved, on average, 90.80%. The SVM, on average, achieved a 90.10% accuracy. Finally, after running the combined model on our separate testing sets, we reached a maximum accuracy of 91.10%. The average accuracies of each of the models is shown in Table II.

We use confusion matrices to analyze the classification performance of each of the classes. In these matrices, the diagonal cells represent the percentage of the test data correctly classified with its true label whereas the other cells represent the percentage of mis-classified data. Figures 3 to 6 show the confusion matrices of our Small CNN, TL_Resnet50, SVM and Combined model.

As shown in confusion matrices, the models had difficulty when it came to handling the differences between hungry baby

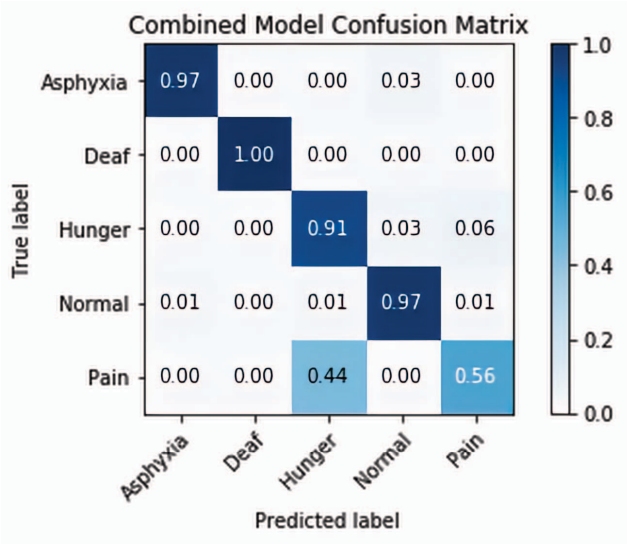


Fig. 6. Confusion Matrix of Combined Model (Ensemble Classification)

cries in comparison to pain baby cries. However, we believe this may be remedied with a more balanced dataset. But, overall, the models performed admirably. The datasets used in most of the related works are different from the one that is used in this work and therefore we do not have a direct way to compare our results..

V. Discussion & Future work

There are many areas that we can modify to improve the reliability of our models. The first being to properly balance our dataset either with more baby cries or at least synthetically using data augmentation. Second, we could try different algorithms to improve our performance. W. Zhong et al. [10] have shown progress with clustering SVMs in protein local structure prediction, and this progress may be transferable to our model. Along with that, Hui Liu et al. [11] have proposed a genetic fuzzy routing algorithm based on fuzzy set theory and evolutionary computing which may be applicable to our models. Additionally, Sara Sabour, Nicholas Frosst, and Dr. Geoffrey E. Hinton's capsule network, which is based on convolutional neural networks may help us classify our spectrograms more effectively. Finally, experimenting with the size and shapes of our spectrogram images may prove to show positive results for our models.

We perform our classification using image spectrogram data which is usually different from the traditional methods chosen to classify audio signals which uses Mel-frequency cepstral coefficients(MFCC) features. In future work, we believe that combining MFCC features along with image data for baby cry classification can further improve the results. We would also like to experiment with various image classification architectures to determine the best model.

An accessible and reliable machine learning approach to baby communication can significantly improve the care and reduce the confusion surrounding infant health. While young

babies may not be able to communicate their needs formally, many papers have proved their cries possibly offer more than enough information to reflect their neurological states and physical needs. Understanding a baby's cry is currently an important topic that if we apply machine learning to, can better human quality of life.

VI. Conclusion

Analysis of baby cries can help in early detection of abnormal conditions in the baby. In our research, we employ machine learning and deep learning models to classify various baby cries taken from the Baby Chillanto Database. We use a transfer learned ResNet50, an SVM, and an Ensemble Classification model of these two previous models to classify spectrograms of baby cries in a multi-class dataset. These models achieve high accuracies, ranging from 90.1% to 91.1%.

Acknowledgment

We'd like to thank Dr. Carlos A. Reyes-Garcia, Dr. Emilio Arch-Tirado and his INR-Mexico group, and Dr. Edgar M. Garcia-Tamayo for their dedication of the collection of the Infant Cry database. We want to express our great gratitude to Dr. Orion Reyes and Dr. Carlos A. Reyes for providing access to the Baby Chillanto database. We would also like to acknowledge Google Colaboratory for providing a free useful development platform that helped us to accelerate our model development process.

References

[1] A. A. of Pediatrics *et al.*, "Early hearing detection and intervention (ehdi)," *Early Hearing Detection and Intervention (EHDI)-aap. org. Accessed August*, vol. 19, 2016.

[2] C. for Disease Control and Prevention, "2016 hearing screening summary — annual data ehdi program — cdc."

[3] O. F. Reyes-Galaviz and C. A. Reyes-Garcia, "A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks," in *9th Conference Speech and Computer*, 2004.

[4] R. Sahak, W. Mansor, Y. Lee, A. M. Yassin, and A. Zabidi, "Orthogonal least square based support vector machine for the classification of infant cry with asphyxia," in *2010 3rd International Conference on Biomedical Engineering and Informatics*, vol. 3. IEEE, 2010, pp. 986–990.

[5] R. Sahak, Y. Lee, W. Mansor, A. Yassin, and A. Zabidi, "Detection of asphyxiated infant cry using support vector machine integrated with principal component analysis," in *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2010, pp. 485–488.

[6] M. Moharir, M. Sachin, R. Nagaraj, M. Samiksha, and S. Rao, "Identification of asphyxia in newborns using gpu for deep learning," in *2017 2nd International Conference for Convergence in Technology (I2CT)*. IEEE, 2017, pp. 236–239.

[7] A. Osmani, M. Hamidi, and A. Chibani, "Platform for assessment and monitoring of infant comfort," in *2017 AAAI Fall Symposium Series*, 2017.

[8] C.-Y. Chang, C.-W. Chang, S. Kathiravan, C. Lin, and S.-T. Chen, "Dag-svm based infant cry classification system using sequential forward floating feature selection," *Multidimensional Systems and Signal Processing*, vol. 28, no. 3, pp. 961–976, 2017.

[9] S. S. eXchange, "Sox - sound exchange," [accessed 2019-07-11]. [Online]. Available: http://sox.ourceforge.net/

[10] W. Zhong, J. He, R. Harrison, P. C. Tai, and Y. Pan, "Clustering support vector machines for protein local structure prediction," *Expert Systems with Applications*, vol. 32, no. 2, pp. 518–526, 2007.

[11] H. Liu, J. Li, Y.-Q. Zhang, and Y. Pan, "An adaptive genetic fuzzy multi-path routing protocol for wireless ad-hoc networks," in *Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Network*. IEEE, 2005, pp. 468–475.