BABY CRY CLASSIFICATIONS USING DEEP LEARNING

Shane Grayson[1] and Wilson Zhu[2]

[1]Windward School, Los Angeles, USA
[2]Diamond Bar High School, Los Angeles, USA

ABSTRACT

New parents are frequently awakened by the cries of their newborn babies. Attempts to stop these cries sometimes result in increasingly louder cries. By first transforming these cries into waveforms, and then into sound spectrograms, the efficiencies and accuracies of different computer learning modules were tested: a support vector machine, a 2-layer neural network, and a long short-term memory model. Finally, an automatic sorter that categorizes each cry was developed. Using this method, it is possible to eliminate error and time wastage when trying to calm a baby. The results of testing the programs demonstrate a high accuracy rate for determining the source of a baby's cries. This program will enable parents to calm their crying babies in a shorter amount of time, giving them more peace of mind, and perhaps allowing them to get more sleep.

KEYWORDS

Infant Cry, Deep Learning, Convolutional Neural Network, Audio Classification.

1. INTRODUCTION

For the module, a combination of convolutional neural networks alongside other programs interpreting audio files was used. Computers did not start with this capability though, so in providing background for the topic, a brief history of the program that can understand written language, natural language processing (NLP), will be given. Natural Language Processing emanates from Noam Chomsky. In his book "Syntactic Structures" published in 1957, Dr. Chomsky theorized that if formatted in his style of grammar, called Phase Structure Grammar, computers would be able to understand human language [1]. In 1958, not long after "Syntactic Structures" was published, one of the first programming language processes, called LISP was released, and capable of giving textual responses in the form of a psychiatric technique called "reflection" [1].

Although high costs stopped most research regarding Natural Language Processing and AI in 1966, by 1980, NLP research was brought back with new, fresh ideas. Researchers abandoned the old ways of mixing statistics with linguistics because that method had not produced a computer program that even came close to carrying out conversations, instead, researchers adopted purely statistical models, even though the rise of early Machine Learning challenged these models; eventually leading to multiple high-level statistical NLP programs being created [1]. Breaking into the modern era, Deep Neural Network Learning has been the leading method for speech recognition, taking advantage of many different architectures, including support vector machines, maximum entropy models, neural networks, and Gaussian mixture models [2].

David C. Wyld: CSITY, NWCOM, SIGPRO, ASOFT, AIFZ, BDIoT, ITCCMA, CLSB, DTMN, MLNLP - 2021
pp. 49-59, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111405

The growing number of architectures are now categorized into three different main groupings: Generative, Discriminative, and Hybrid deep learning architectures. Discriminative architectures can use visible data to determine patterns and characterize different aspects of the data [2]. Convolutional Neural Networks are just one of many examples of architecture in discriminative classification. These modules consist of a convolutional layer and a pooling layer, usually stacked, in order to form deep models [2]. They can enable computer programs to not only understand the language but to also understand the sentiment humans put behind that language. These models have been used in the healthcare field as NLP's which can deduce underlying meanings in forms filled out by patients to a high degree of accuracy, with similar methods being developed to be used on social media [3].

1.1. Existing Methods

These models can be applied in a broad array of situations, so it was difficult to narrow down all of the possible problems to the one that was settled on. Interest in developing an automated lab cry categorizer sprouted from the want to reduce the anxiety and sleep deprivation caused when new parents are forced to wake up at all hours of the night and care for their babies. Parents frequently have to test a multitude of ideas before one finally works. The effects of constantly trying to comfort a child are numerous and range from interrupting a parent's circadian rhythm (which, in particular, intensifies a mother's fatigue after the birthing process), to anxiety at the thought of harming one's child, and even depression [4]. While the added interaction may increase familiarity and trust in the parent, these side effects counteract that and can damage the parent-child relationship [4].

For these reasons, allocating the stress and work put on parents is a high priority. While there have been no groups using the same methods combined with this idea, there have been groups with a similar idea and separate execution, with their own pros and cons. For example, most recently a group participating in the 2019 10th International Conference on Computing, Communication and Networking Technologies tackled this issue by using Statistical Feature Extraction and Gaussian Mixture Models [5]. This team achieved an accuracy of 81.27%.

1.2. Method Used

As mentioned above, the most recent idea used Statistical Feature Extraction and Gaussian Mixture Models and was completed with an accuracy of 81.27% identifying a total of five different reasons for crying [5]. Gaussian Mixture Models are probability density functions that represent the weighted sums of component densities of a Gaussian [6]. While the use of Gaussian Mixture Models is a fascinating solution for determining differences between audio clips, and for inferring meaning through the different components of these cries, this method falls short on accuracy compared to other methods. That said, it does make up for this decreased accuracy by being extremely resistant to overfitting, allowing immense amounts of training to be done. Moreover, higher accuracy can be achieved using a support vector machine, while continuing to avoid the problem of overfitting. Overall, however, Gaussian Mixture Models are not the optimal method to use.

The methods used in this paper improves upon the accuracy used in previous approaches centering around Gaussian Mixture Models while avoiding issues like overfitting. This was achieved by first labeling the data into the reason for their cries: "belly_pain", "discomfort", "tired", "hungry", and "burping". After labeling all of the data, two built-in tensor flow tools were used to first decode the audio files and then transform them into waveforms, while simultaneously labeling them. This makes each audio file visible as a graph where some patterns are able to be seen; for instance, a pattern observable in the "hungry" waveforms is continuous

and consistent pulse durations, which is seen as well in the “discomfort” labeled waveforms, just with much larger average crest and trough amplitudes. These both differ from the “tired” group, however, whose amplitude starts small comparatively and decreases alongside the wavelength as the cry continues. While the trends displayed by one group are difficult to notice, they are, in fact, discernible by the human eye. In order to enlarge these changes and highlight the trends previously seen, it was necessary to convert these waveforms into sound spectrograms on a logarithmic scale with labeled axes time and frequency.

Now, with the trends obvious, the three different models were trained, and the results were compared to determine the best way of interpreting this data. A 2-layer neural network, support vector machine, and long short-term memory model with results similar inaccuracy, but a loss rate much better from the support vector machine were used, leading to the conclusion that this method is best when categorizing baby cries by meaning.

1.3. Evaluations

Through multiple different applications, including a 2-layer neural network, support vector machine, and long short-term memory model, it can be concluded that the support vector machine is the most accurate method in determining the cause of a baby crying. Finishing with an accuracy of 86% compared to the other accuracies of 81%, it can be concluded that of the methods tested, the support vector machine is superior in classifying the sentiment behind a baby’s cries.

1.4. Paper Structure

The rest of this paper is organized in the following manner: Section 2 details the process and solution used to address the problem offered. Section 3 gives an overview of the results recorded during the evaluation of each program on the validation set as well as an analysis of the meaning behind these results. Finally, Section 4 offers a conclusion as well as a potential improvement upon the execution and ideas in future work.

2. OVERVIEW OF OUR APPROACH

In order to construct the processes described below, both Keras and TensorFlow were used as they comprise the largest python programming platform. Despite not being the most user-friendly software, TensorFlow allows for many different levels of customization, enabling the use of different methods. The model is split into three major processes: representing each data file as a waveform, transforming this waveform into a sound spectrogram, and finally using a series of different modules to train and test on these spectrograms.

The data files of babies crying were sourced from GitHub, where volunteer participants would download the Donate-a-cry application for either iOS or Android and submit an audio file of a baby crying along with the reason for why they were crying. These cleaned and filtered files are the ones used from GitHub as data samples for this paper [7]. Each file, when imported, was given a name matching the reason identified by the volunteers as the reason for crying: “belly_pain”, “discomfort”, “tired”, “hungry”, “burping”.

There was a total of 457 samples used. Of these 457 samples 360 were used for training, 40 for validation, and the other 57 were reserved for testing; this final accuracy is what was used to determine the overall effectiveness of the program. To transform these files into waveforms, they were first decoded using a TensorFlow built-in program. Once decoded the values here graphed

with the y-axis representing time and x-axis representing amplitude. Originally the values for amplitude were given from -32768 to 32767, but were shrunk to fit between -1.2 and 1.2; once completed they looked as shown in Figure 1:

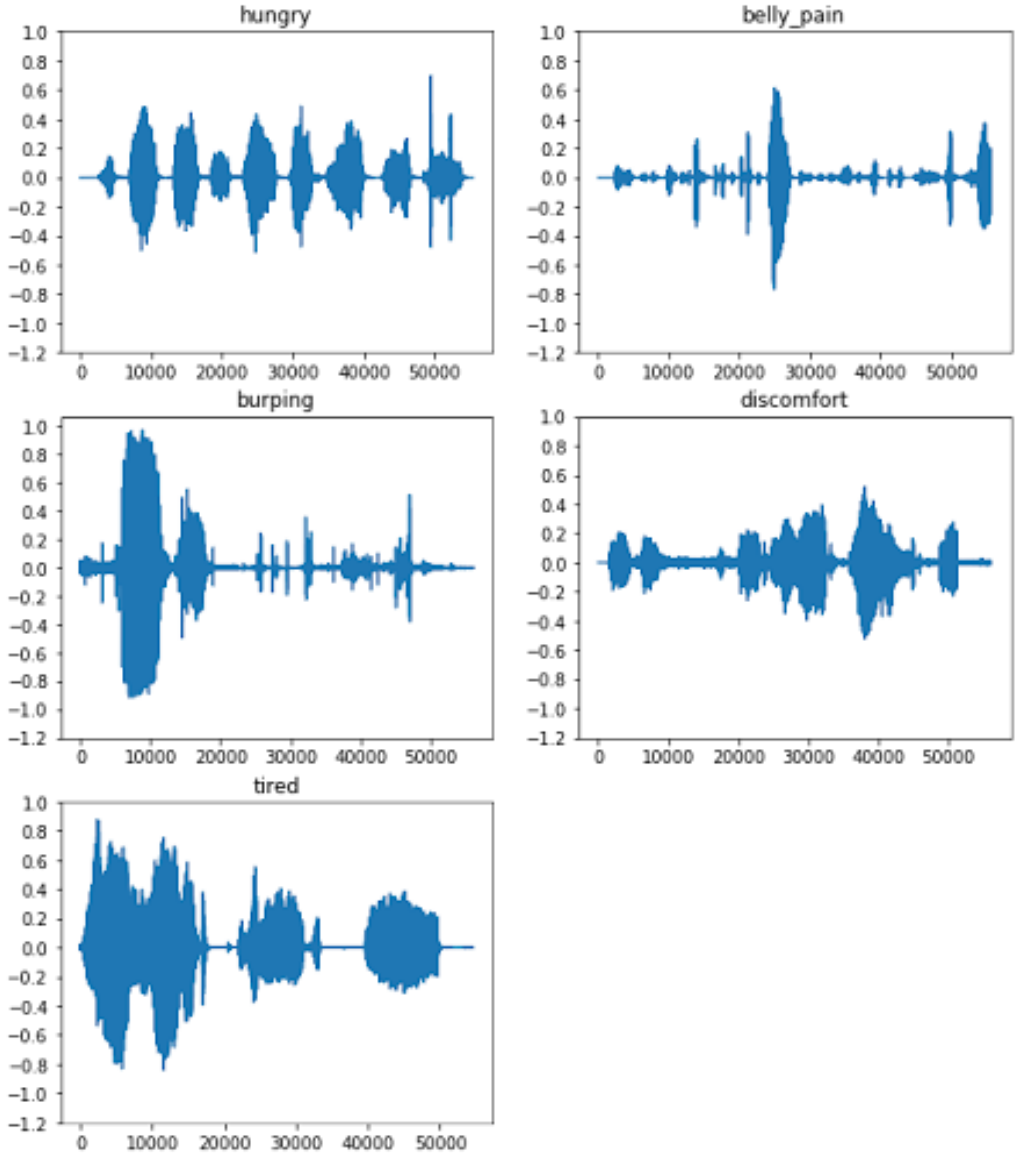

Figure 1. Examples of Data Waveforms

Afterward, each waveform of fewer than 60,000 samples was padded to be this length. They were then converted into spectrograms where the y-axis was time ranging from 0-60,000, the x-axis was measured in frequency from 0-120, and color represents amplitude with lighter color showing higher amplitudes and darker color showing lower amplitudes. To help accomplish this, the Fourier Transform mathematical concept has been used, which transforms the data from an audio signal into a frequency domain [8]. Figure 2 shows an example of these results.

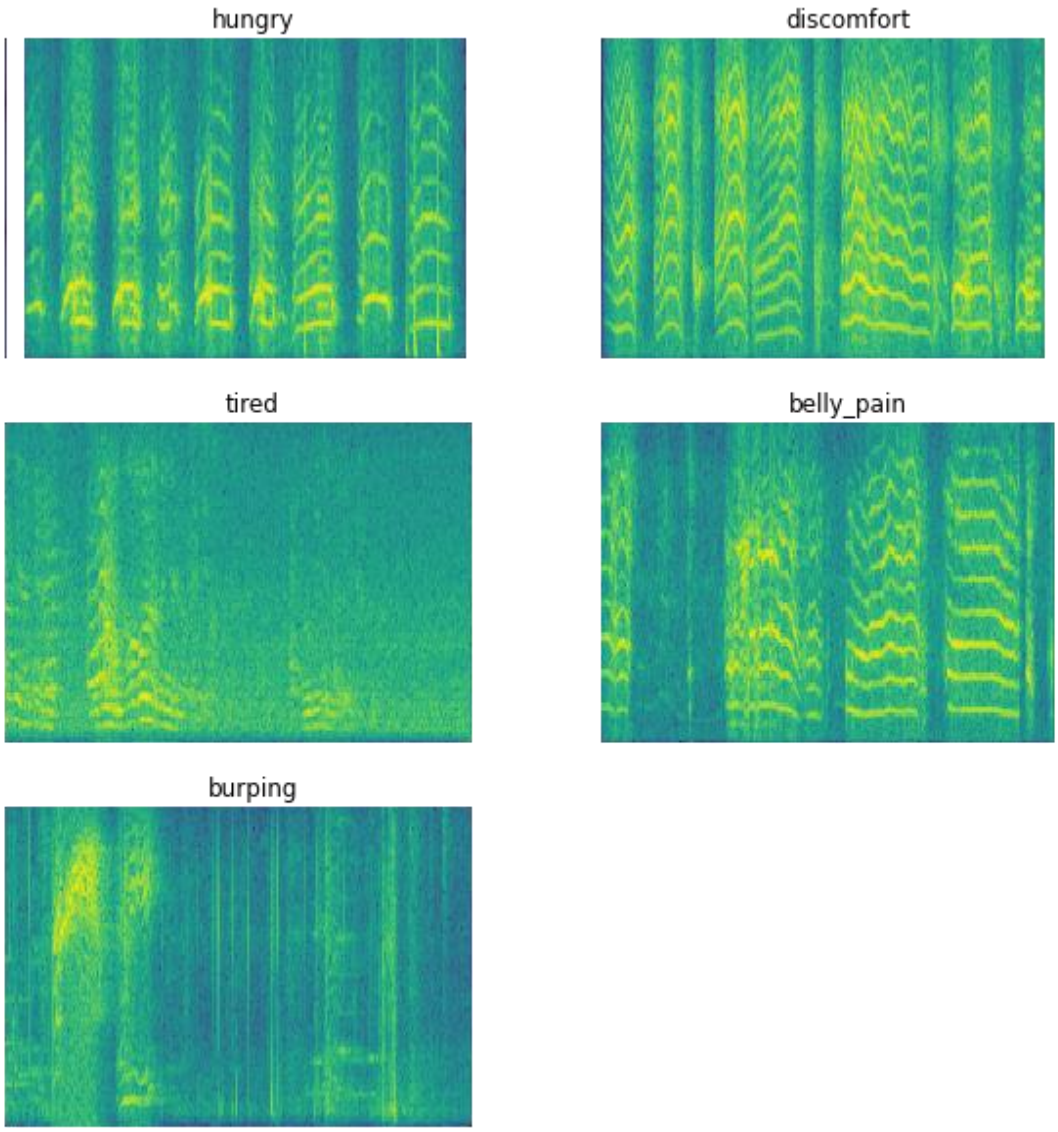

Figure 2. Example of Data Spectrograms

There is then a convolutional neural network created which starts by resizing the data. From this point on there are differences in the code for the three different testing methods: one for a 2-layer neural network, the next for a support vector machine, and finally a long short-term memory module. For the 2-layer neural network, the images of the spectrograms are first resized, then each pixel is normalized through a normal layer. Next, there are two convolutional layers. They are used because they can extract patterns and features from the pixels. Convolutional layers use a filter or a set of weights that, when multiplied by the input, showcases the probability of that patch of pixels representing a feature the filter is trying to find. This filter scans the image before moving a certain number of pixels and scanning once again. That movement is called a “stride”. The output of one of these layers is a smaller image with more showcased patterns.

Both of the above-described layers, as well as other convolutional layers, use the Rectified Linear Unit (Relu). Deep learning modules that use a gradient algorithm tend to get trapped at a local minimum, in order to avoid this, the Relu function is used which speeds up the convergence learning of the module [9]. Following those is a pooling layer, which reduces the image even

further by taking the largest value in an area and passing only that value forward to the next level. Later in this sequential model, two dense layers are used. Dense layers are layers in which every neuron in the previous layer is connected and sends a signal to each neuron in the dense layer.

The SVM (Support Vector Machine) model starts similarly with a resizing and normal layer. There are then two convolutional layers and a pooling layer. Afterward, Random Fourier Features using the gaussian radial basis function distributing parameters maps out, from the input layer's dimensions to lower dimensions to create a randomized feature space based on the approximate shift-invariant kernels. The SVM model is finished with a final dense layer.

The last model is an LSTM (Long Short-Term Memory) model also starting with resizing and a normal layer. The images are then reshaped, and a convolutional LSTM layer is applied. When doing this the image of a spectrogram is converted into a sequence of parts of images that are then passed to a convolutional layer before being passed to an LSTM which spots trends in this sequence and predicts the label. The learning rate of these models can be seen represented in Figure 3. This is used to stop the model training at a global minimum rather than a local one.

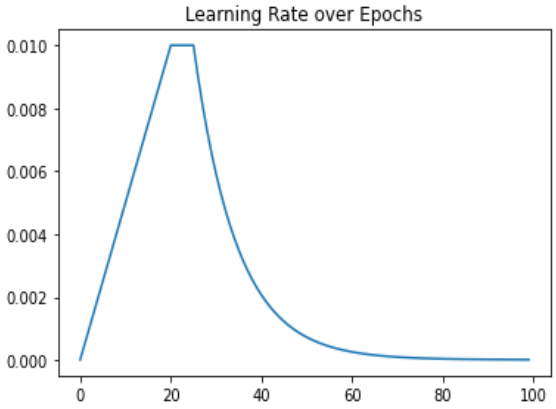

Figure 3. Learning Rate of Models over Epochs

3. EXPERIMENT

To determine which model was best in this situation, the accuracies of each were compared as shown in the below figure over a total of 100 epochs. For proper results, the three programs were all executed on Google Colab with the hardware accelerator being a GPU. It can be seen in Figure 4 that the SVM analyzed the data with higher accuracy of 86% when compared to the 2-layer neural network and LSTM with 81%.

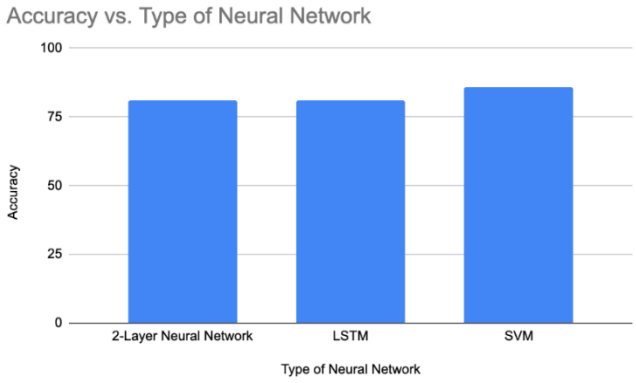

Figure 4. Accuracies of the Three Models for the Testing Set

For each of the three models, the sparse categorical cross-entropy loss function from logits is used when there are more than two label types. The SVM reported much lower levels of loss with just 1.0053 when compared to the other two models as the 2-layer neural network reported 3.53035 and the LSTM, the greatest of the three, with 3.8031. These results can be seen in Figures 5, 6, and 7. The accuracies of all three on the validation sets can also be seen in the graph with the SVM having an accuracy of 77.5%, the 2-Layer Neural Network with 75%, and the LSTM model with 72.5%.

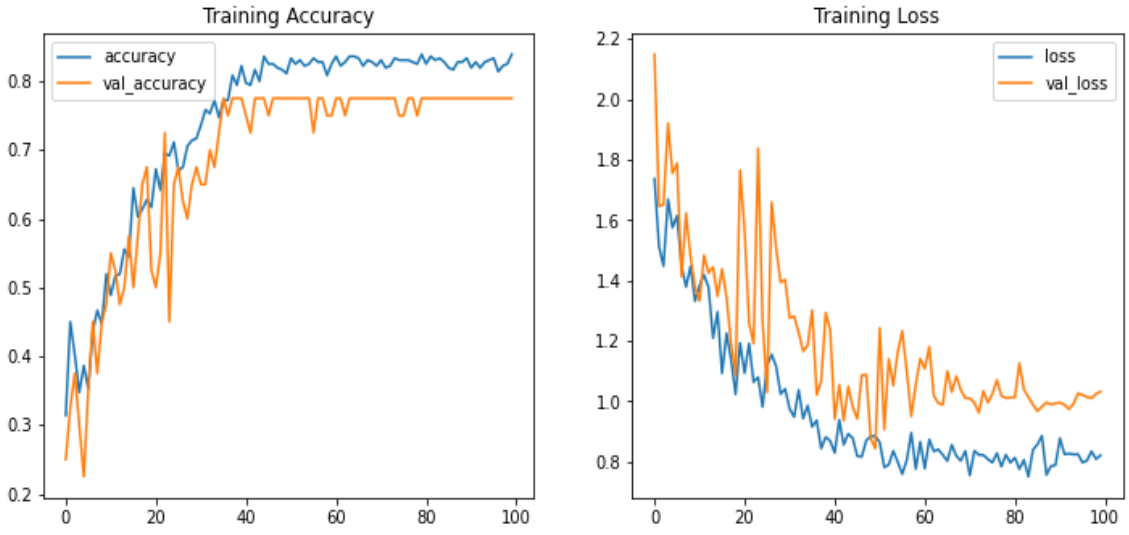

Figure 5. Accuracy and Loss Over Epochs of the SVM for Training and Validation Sets

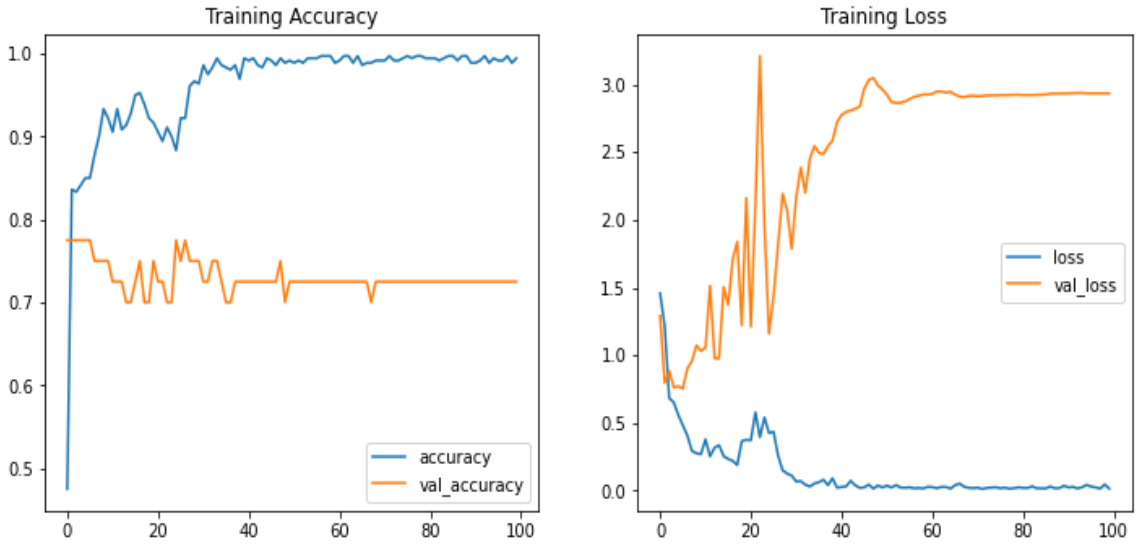

Figure 6. Accuracy and Loss Over Epochs of the 2-Layer Neural Network for Training and Validation Sets

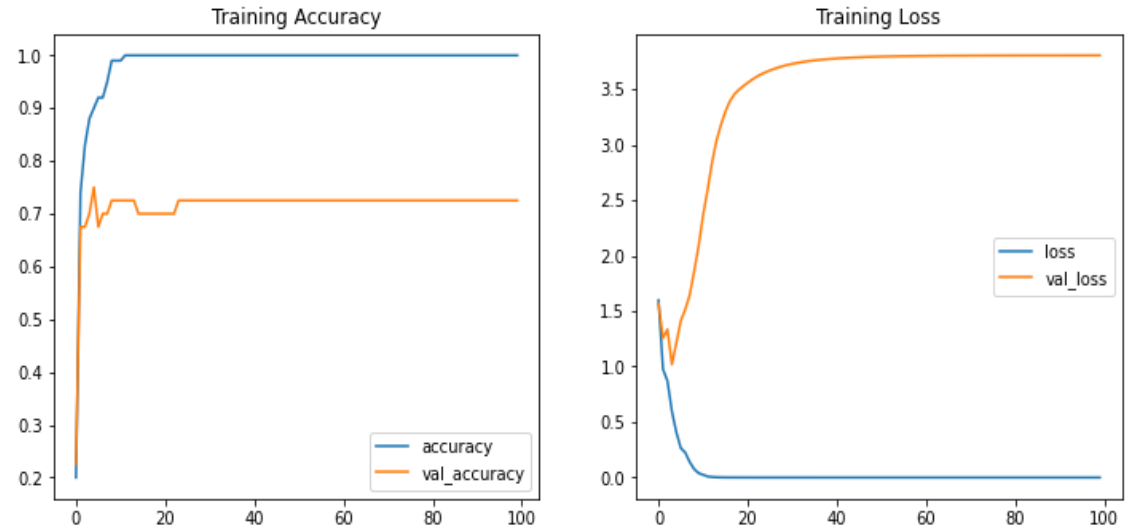

Figure 7. Accuracy and Loss Over Epochs of the LSTM for Training and Validation Sets

3.1. Analysis

Through the following analysis, it can be concluded the model most fit for determining the cause of a baby's cries is the SVM, which supports the highest accuracy without signs of severe overfitting. The relatively high accuracy of the SVM can be attributed to the high number of parameters as well as the small size of data. SVM's can outperform neural networks when the data size is low, so this is one possible reason for the neural networks' relatively low accuracy. SVM's perform at such high levels in small data sets because they create a hyperplane separating the observations they make [9]. This is useful because the SVM will always find a solution when classifying by increasing the dimensionality of its hyperplane to a level where there is a solution [10]. Another reason for the higher accuracy achieved when using the SVM is the number of layers. Given that if the number of parameters in the 2-layer neural network was increased to

match that of the SVM, the 2-layer neural network should perform at a higher level of complexity, offering a much-improved accuracy.

As for the LSTM, there is another reason why it performed poorly. LSTM specializes in predicting the next picture in a sequence of pictures, or in a video, so it is out of its depth trying to spot trends and assign those trends to labels. To combat this each image of a spectrogram was split into parts of the whole and used as parts in a sequence in order to imitate a video, but to no benefit beyond what the other models would provide. Another possible hindrance to the LSTM model could be the lack of a second convolutional layer. Both of the other two models contained two convolutional layers compared to the LSTM model's one. Without this second layer, pulling and analyzing patterns and trends is made more difficult, potentially leading to a lower accuracy once the model is required to identify trends. While the validation accuracy is lower than that of the SVM's, the testing accuracy does reach perfection. This, along with the loss function graph, indicates severe overfitting. One explanation for this is that LSTM models have the ability to hold memory, so overtraining on a small size of data will lead to the model memorizing the data rather than learning the trends that each piece demonstrates.

Overfitting can also be seen in the case of the neural network. This can be seen in Figure 6 with the loss being much higher in the validation set than in the training set. While, as stated earlier, an increased number of parameters can be the catalyst that allows for increased accuracy and complexity, a high number of trainable parameters can lead to overfitting, which is possibly what is being seen here. The number of trainable parameters possessed by the 2-layer neural network is over ten times larger than either of the other two modules. Besides the SVM's tendency to thrive in small data size experiments, the number of trainable parameters could be a reason for not overfitting, in contrast to the 2-layer neural network which contains twenty times the number of trainable parameters.

4. RELATED WORKS

Other methods have been used in combination with a similar premise. One such example is by a group from Yunlin University who used a combination of convolutional neural networks to determine if the spectrogram fed to it was a baby's cry and then classify that cry's cause as one of four reasons [11]. Another group Universitas Indonesia paired convolutional neural networks with recurrent neural networks, allowing a more streamlined process in which the recurrent neural network learns off of the features extracted by the convolutional neural network [12]. In the same vein, a group from Georgia State University focused on convolutional neural networks as well, trying to improve upon them by using a multi-stage convolutional neural network with a hybrid feature set and prior knowledge [13].

In opposition to these ideas, a group from Koç University used a capsule network in direct comparison to regular convolutional neural networks and fed the audio signals from spectrograms created by their audio file data into the capsule network [14].

The first three models all use convolutional neural networks as a central part of their model, while the second and third ones try to improve upon the normal convolutional network by introducing new features, but still in hopes that a convolutional neural network is best suited for this task. The fourth group takes a different approach. Instead, this group compares a capsule network to convolutional neural networks in support of their model.

5. CONCLUSION

In order to solve the problem of determining the cause behind a baby's cry, first, a dataset with labeled reasons for why babies were crying was found. From here each file was assigned a label before being transformed into a waveform so that it could be seen and interpreted by programs. Waveform trends were also barely noticeable to the human eye inside of one label. In order to make these trends more obvious, each file was again transformed, now as a spectrogram. The spectrogram adds another layer that can be analyzed, making the trends more obvious to the programs. From here, three different programs were trained and tested on this data: a support vector machine, a 2-layer neural network, and a long short-term memory model (all of which incorporated at least one convolutional layer). After testing was complete, due to the highest accuracy and lowest loss, it was determined that for this experiment, the SVM was best suited to label the causes for babies to cry.

5.1. Current Limitations

There were some shortcomings to this experiment. The most glaring is the data size which disproportionately negatively affected the 2-layer neural network and long short-term model. Another place for improvement would be the number of reasons for which the baby is crying because they are currently limited to five. In practicality, there are a multitude of reasons that cause a baby to cry and limiting that number to five while training will lead to a much lower accuracy when being tested in real-life situations.

5.2. Future Works

In the future in order to improve upon this research, it would be beneficial to have access to a much larger data size to fairly test each module and accurately determine which model is optimal for this task. Finally, adding on as many causes to the cries is another step in improving this paper, which will significantly improve the accuracy when the testing is performed on actual babies.

REFERENCES

[1] Foote, Keith D. "A Brief History of Natural Language Processing (Nlp)." DATAVERSITY, 17 June 2019, www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/.

[2] Deng, Li. *Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey*. 2012, www-edlab.cs.umass.edu/cs697l/readings/Three%20Classes%20of%20Deep%20Learning%20Architectures.pdf.

[3] Rajput, Adil. "Natural Language Processing, Sentiment Analysis, and Clinical Analytics." Innovation in Health Informatics, Academic Press, 15 Nov. 2019, www.sciencedirect.com/science/article/pii/B9780128190432000034.

[4] Kurth, Elisabeth, et al. "Crying Babies, Tired Mothers: What Do We Know? A Systematic Review." Midwifery, Churchill Livingstone, 20 Sept. 2009, www.sciencedirect.com/science/article/abs/pii/S0266613809000692.

[5] K. Sharma, C. Gupta and S. Gupta, "Infant Weeping Calls Decoder using Statistical Feature Extraction and Gaussian Mixture Models," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944527.

[6] Reynolds, Douglas. *Gaussian Mixture Models*. 2009, leap.ee.iisc.ac.in/sriram/teaching/MLSP_16/refs/GMM_Tutorial_Reynolds.pdf.

[7] Gveres. “Gveres/Donateacry-Corpus: An Infant Cry Audio Corpus That's Being Built through the Donate-a-Cry Campaign - See Http://Donateacry.com.” GitHub, github.com/gveres/donateacry-corpus.

[8] Chaudhary, Kartik. “Understanding Audio Data, Fourier TRANSFORM, FFT, Spectrogram and Speech Recognition.” *Medium*, Towards Data Science, 18 Jan. 2020, towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520.

[9] K. Hara, D. Saito and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1-8, doi: 10.1109/IJCNN.2015.7280578.

[10] Luca, Gabriele De. “SVM vs Neural Network.” Baeldung on Computer Science, 9 Sept. 2020, www.baeldung.com/cs/svm-vs-neural-network.

[11] Chang CY., Tsai LY. (2019) A CNN-Based Method for Infant Cry Detection and Recognition. In: Barolli L., Takizawa M., Xhafa F., Enokido T. (eds) Web, Artificial Intelligence and Network Applications. WAINA 2019. Advances in Intelligent Systems and Computing, vol 927. Springer, Cham. https://doi.org/10.1007/978-3-030-15035-8_76

[12] Maghfira1, Tusty Nadia, et al. “Iopscience.” *Journal of Physics: Conference Series*, IOP Publishing, 1 Apr. 2020, iopscience.iop.org/article/10.1088/1742-6596/1528/1/012019.

[13] Ji C., Basodi S., Xiao X., Pan Y. (2020) Infant Sound Classification on Multi-stage CNNs with Hybrid Features and Prior Knowledge. In: Xu R., De W., Zhong W., Tian L., Bai Y., Zhang LJ. (eds) Artificial Intelligence and Mobile Services – AIMS 2020. AIMS 2020. Lecture Notes in Computer Science, vol 12401. Springer, Cham. https://doi.org/10.1007/978-3-030-59605-7_

[14] Tuğ̆tekinTuran, and ErzinEngin. *Monitoring Infant’s Emotional Cry in Domestic Environments Using the Capsule Network Architecture*. Sept. 2018