JOURNAL OF
EMERGING INVESTIGATORS

Article

# Using neural networks to detect and categorize sounds

**Breanna Micciche[1], Terese Grateful[1]**
[1] Council Rock High School North, Newtown, Pennsylvania

## SUMMARY

**Artificial neural networks with accurate noise detection can help people with hearing loss be aware of important sounds. Neural networks have already been used for medical purposes and noise detection. However, they typically require large amounts of training data, and background noise can severely decrease the accuracy of the audio detection. The purpose of this project was to examine whether a feed forward neural network (FFNN), recurrent neural network, or convolutional neural network is most effective at audio classification. All three neural networks were trained using the same data of bell sounds, knocking sounds, guitar sounds, and talking. We hypothesized that the convolutional neural network would be the most accurate because it is structured to use more data when it makes predictions and that the FFNN would be the quickest because it requires the least amount of calculations to make predictions. The accuracy of the neural networks was tested with new randomly selected audio of the four categories with no background noise, white noise, environment noise, and busy background noise. Results were compared with the accuracy and times of human participants listening to and categorizing the same sounds. The convolutional neural network was the overall most accurate of the three neural networks, but the feed forward neural network was more accurate when there was little background noise. The recurrent neural network was the least accurate. The feed forward neural network was the fastest among the neural networks and the participants.**

## INTRODUCTION

Audio detection with machine learning has several uses for assisting people, such as alerting people with hearing loss to sounds that are considered important, like fire alarms (1). This technology is also used to monitor and detect medical issues in people using heartbeat and breathing patterns (2).

Artificial neural networks are used in machine learning to detect patterns by taking inputs and performing computations to produce an output (3). Neural networks, particularly feed forward (FFNN), recurrent (RNN), and convolutional (CNN) neural networks, are common methods of audio categorization. Neural networks use training data to learn patterns and improve the accuracy of categorizing information (3). The number of samples used for training can be increased or decreased to adjust the amount of data used for the neural networks to detect patterns. There is not a set number of samples used, however, more samples tend to produce better results (4).

An artificial neural network is made of layers of nodes (5). The input is the first layer and is the data provided to the neural network (5). The output is the last layer and is the data that is produced by the neural network (5). In between these two layers are hidden layers, where the computations are done to the input data to produce the output (5). These nodes are connected by weights, which is the importance of the node (5).

FFNNs are a simple neural network compared to RNNs and CNNs because the connections in the network do not form a cycle, and information only gets processed in one direction. In an FFNN, the inputs are multiplied by weights in the hidden layer and the sum of those values is the output. The output is compared to intended values to classify the input and train the neural network, and then the weights are adjusted during each iteration, which is called an epoch (6). RNNs are similar to FFNNs, but instead of only going through the hidden layer once, the values from the hidden layer go back into the same or previous layers (7). CNNs specialize in data with grid-like topologies, such as images, and can be applied to audio when the audio analysis graphs are converted into images. CNNs typically have convolutional layers and pooling layers. The convolutional layer detects patterns in different subregions of the input. The pooling layer reduces the size of the input while keeping important structural data to reduce the required computing power (8).

Other studies have compared the accuracy of different neural networks, such as CNNs and FFNNs (9,10). In particular, work to detect properties in images of leaves or of X-rays found that CNNs were significantly more accurate than FFNNs. These studies demonstrate the potential uses of neural networks for categorization.

In this study, a FFNN, RNN, and CNN were trained to categorize sounds, specifically a bell, a guitar, talking, and knocking, in different types of audio backgrounds. Since neural networks require numerical data, the audio was converted into mel spectrograms. The four different background types used were no background noise, white noise, environment noise, and busy noise (11-13). The environment noise featured sounds that could be heard outdoors, like wind blowing and birds chirping. The busy noise featured sounds that could be heard in a busy store, like objects being moved, people moving, and people talking. Mel spectrograms are graphical representations of audio frequencies over time (14). The images of the mel spectrograms were then converted into two-dimensional arrays of the RGB values of the pixels. Additionally, the training sample sizes of the neural networks were decreased to examine the potential effect of sample size

https://doi.org/10.59720/23-214

on accuracy.

As a control, we tested the ability of human participants to categorize the same sounds in the same types of audio backgrounds. We hypothesized that the CNN would be the most accurate at categorizing the sounds because of the use of subregions in the convolution layers to make predictions, and the FFNN would be the quickest because it requires the fewest calculations, making it less resource-demanding. While the CNN was the overall most accurate out of the neural networks, it was less accurate than the human participants. It was also less accurate than FFNN when there was no background noise. The FFNN was the fastest, but there was very little difference between the speed of the FFNN, the RNN, and the CNN.

## RESULTS

The neural networks were tested by using each of them to predict the category of a total of 100 randomly selected audio files of knocking, talking, guitar, or bell sounds for each background type. The time it took for the neural networks to make the predictions, the prediction they made, and the correct category were recorded. This data was used to calculate the average time, the overall accuracy of the neural networks, and the accuracy of each neural network for each background type. We generated mel spectrograms of each recording, which varied by background types (**Figure 1**).

The human participants were tested by having them take a reaction time test and then having them categorize five random sounds for each background type. Reaction time, time to categorize, chosen category, and correct category were all recorded. We used the defined categories to calculate the average time it took for the participants to categorize the sounds, the overall accuracy, and the accuracy in each category.

We found that the human participants were more accurate than all the neural networks with every background type, except when there was no background noise (**Figure 2**). Overall, the CNN was the most accurate of the neural networks; however, with quieter backgrounds, like no background noise or environment noise, the FFNN was the most accurate. The white noise background had the greatest difference in accuracy between neural networks and human participants, a 60% difference. The neural networks had

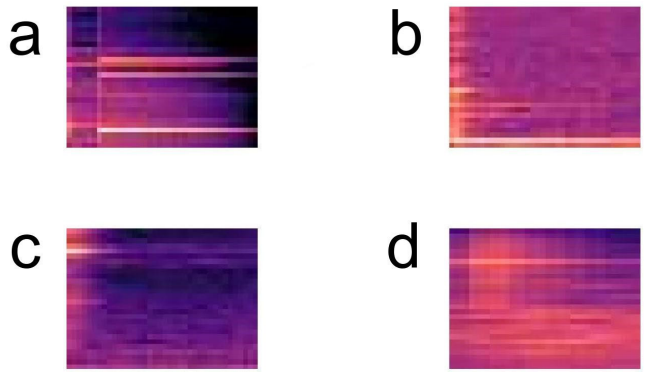


**Figure 1: Mel spectrogram of a bell sound in the different backgrounds.** Mel spectrogram of a bell sound over one second with (a) no background noise, (b) white noise, (c) environment noise, and (d) busy background noise. A lighter color means a higher volume in decibels.

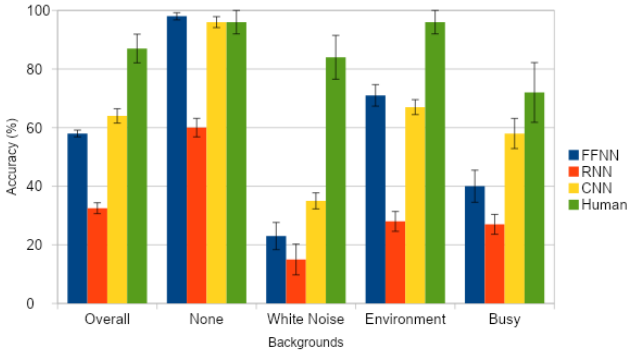


**Figure 2: Accuracy of the neural networks and participants for the different backgrounds.** Percent accuracy of FFNN (n=5), RNN (n=5), CNN (n=5), and Human Subjects (n=5) categorizing bell sounds, knocking sounds, talking, and guitar with no background noise, white noise, environment background noise, and busy background noise. The data for the neural networks were gathered by repeatedly testing the models with 100 mixed sounds for each of the different background types, and the data from the human subjects was gathered from testing them with a total of 5 mixed sounds per participant for each of the 4 backgrounds. Error bars represent standard deviation. Analysis with two-way ANOVA of neural network type, background noise and accuracy revealed statistical significance with $p < 0.001$. Results of Tukey test comparisons shown in Table 1.

the greatest difference in accuracy compared to the human participants in the white noise background (24% accuracy for the neural networks compared to an accuracy of 84% for the human participants). The human participants had the lowest accuracy in the busy background (72% compared to an overall accuracy of 87%) (**Figure 2**).

We observed that background noise had a significant effect on the accuracy of a neural network (two-way ANOVA, $p < 0.001$, **Figure 2**). We used the Tukey method to compare each pair of background types (**Table 1**).

The neural networks each identified the sounds in less than 0.065 seconds, while the human participants all required more than 1.5 seconds to identify the sounds (**Figure 3**). We observed that the type of neural network had a significant effect on the speed of the neural networks (two-way ANOVA, $p < 0.05$, **Figure 3A**). The Tukey method did not identify any statistically different pairs of background types (**Table 2**).

The human participants all had times greater than 1.5 seconds, with a maximum of 3.4 seconds. We observed that the background type did not affect the time for the human participants to categorize each sound (one-way ANOVA, $p = 0.142$, **Figure 3B**).

To optimize the neural networks for future usage, the FFNN and CNN were tested with different training sample sizes. However, the RNN was not tested with different sample sizes because the accuracy was much lower than the other neural networks in all categories (**Figure 2**).

The accuracy of the FFNN increased with sample size and peaked at 75 samples (**Figure 4A**). We observed that the sample size had a significant effect on the accuracy of the FFNN (two-way ANOVA, $p < 0.001$, **Figure 4A**). The CNN also had an increased accuracy with increased sample size; however, it had a maximum accuracy at 100 samples, and the range of the overall accuracy is greater than the FFNN (22% range of accuracy for the FFNN compared to 27% for the CNN) (**Figure 4**). We observed that the sample size had

https://doi.org/10.59720/23-214

| | Overall | None | White Noise | Environment | Busy |
|---|---|---|---|---|---|
| **Overall** | x | $p < 0.05$ | $p < 0.05$ | $p > 0.05$ | $p < 0.05$ |
| **None** | $p < 0.05$ | x | $p < 0.05$ | $p < 0.05$ | $p < 0.05$ |
| **White Noise** | $p < 0.05$ | $p < 0.05$ | x | $p < 0.05$ | $p < 0.05$ |
| **Environment** | $p > 0.05$ | $p < 0.05$ | $p < 0.05$ | x | $p < 0.05$ |
| **Busy** | $p < 0.05$ | $p < 0.05$ | $p < 0.05$ | $p < 0.05$ | x |

**Table 1: Statistically significant pairs of background types for accuracy.** Table with the pairs of background noise types and whether the difference in their categorization accuracy is statistically significant.

a significant effect on the accuracy of the CNN (two-way ANOVA, $p < 0.001$, **Figure 4B**). We used the Tukey method to compare each amount of training samples used for the FFNN and CNN (**Table 3, 4**).

## DISCUSSION

The CNN was the most accurate overall of the three neural networks. However, the FFNN was more accurate with no background noise and with the environment noise, which were the two quietest backgrounds. Due to the lower interference from the backgrounds, the mel spectrograms for those two backgrounds were most similar to the ones used to train the neural networks, so the FFNN was able to categorize them very accurately. For those two background types, the FFNN and CNN performed similarly, so with a larger training sample size, there would be more data for the neural networks to find patterns. So, the differences in the accuracy could change, and one neural network may be better than the other, supporting future research. The CNN outperformed the other two neural networks in the white noise and busy background types, which had background audio at a similar volume to the primary sounds. The CNN was likely most accurate with these because it is designed to filter out unimportant data and consider the surrounding data during each calculation, making it better prepared to deal with the stronger background noise. This finding also supports the idea that the CNN would be the best type of neural network to use for real-world noise detection because it can better handle background noise.

The accuracy of both the participants and neural networks decreased when there was stronger background noise. As previously mentioned, the accuracy of the neural networks likely decreased because the additional noise made it more difficult to detect the patterns on which they were trained. However, the accuracy of the participants dropped to 72% when using the busy background, which could be due to two different reasons. First, the busy background was similar to the sounds of a busy shop, including people talking in the background and making noises while moving around. These could have easily been mistaken by both neural networks and participants as the talking and knocking sounds that they were trying to categorize. Second, the additional auditory information could have distracted the participants, making it more difficult for them to notice the primary sounds they were trained to hear. Since these neural networks are meant to perform similarly to a human, the human participants' accuracy dropping when there is more background noise provides a benchmark for how well a neural network should perform.

The FFNN was the fastest out of the neural networks and participants by a small amount and was statistically significant. The neural networks, in general, were much faster than the participants because computers can process information much faster than humans can. The FFNN was likely the fastest because it is the most computationally simple type of neural network, so fewer calculations are required. All neural networks had a simple structure with no drop-out layers, which prevent the neural network from detecting false patterns by dropping nodes, and they all used the same compiler settings (15).

Both the FFNN and CNN increased in accuracy with increasing sample size. Although they were in a similar range of accuracy for 5 and 100 samples for all the background

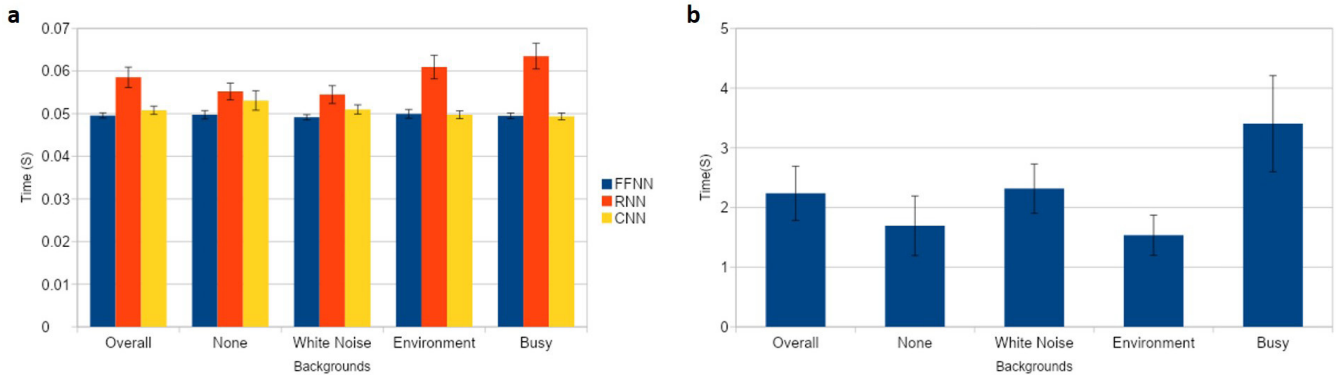


**Figure 3: Average neural network times for the different sound backgrounds. A)** Average time in seconds of FFNN (n=5), RNN (n=5), and CNN (n=5) at categorizing bell sounds, knocking sounds, talking, and guitar with no background noise, white noise, environment background noise, and busy background noise. The data for the neural networks was gathered by repeatedly testing the models for the different background types and finding the change in time from before and predicting each sound category. Average time in seconds of human participants at categorizing bell sounds, knocking sounds, talking, and guitar with no background noise, white noise, environment background noise, and busy background noise. Data shown as mean ± SD (n=5). Error bars represent standard deviation. Analysis with two-way ANOVA of neural network type, background noise and time revealed statistical significance with $p < 0.05$. Results of Tukey test comparisons shown in Table A2. **B)** Average time in seconds of human participants at categorizing bell sounds, knocking sounds, talking, and guitar with no background noise, white noise, environment background noise, and busy background noise. Data shown as mean ± SD (n=5). Error bars represent standard deviation. Analysis with two-way ANOVA of background noise and human participant time revealed not statistically significant, with $p = 0.143$.

https://doi.org/10.59720/23-214

| | Overall | None | White Noise | Environment | Busy |
|---|---|---|---|---|---|
| **Overall** | x | p > 0.05 | p > 0.05 | p > 0.05 | p > 0.05 |
| **None** | p > 0.05 | x | p > 0.05 | p > 0.05 | p > 0.05 |
| **White Noise** | p > 0.05 | p > 0.05 | x | p > 0.05 | p > 0.05 |
| **Environment** | p > 0.05 | p > 0.05 | p > 0.05 | x | p > 0.05 |
| **Busy** | p > 0.05 | p > 0.05 | p > 0.05 | p > 0.05 | x |

**Table 2: Statistically significant pairs of background types for categorization speed.** Table with the pairs of background noise types and whether the difference in their categorization speed is statistically significant.

types, the FFNN had a smaller accuracy range over the different sample sizes compared to the CNN, making it more consistent, so more training samples are not necessary. The accuracy range is important because it would mean a user would not need to record many samples. Since both the CNN and the FFNN had statistically significant differences between the accuracy of five samples and almost every other sample size, five samples are not enough data to reliably detect accurate patterns in data. The lack of a significant difference between 50 and 100 training samples may suggest that increasing the amount of training data beyond this point would have diminishing returns. Potentially, with further modification and optimization of the neural networks, the amount of training data could be further reduced.

Due to the large number of sounds (400 sounds) that were needed to train the neural network, they were all either recorded by the researchers or found online. Although this made it easy to access many types of sound, there was not a consistent quality or volume, which could have had a negative impact on the training process due to the variation. Also, since the neural networks were only trained on sounds without background noise, the neural networks were more accurate at predicting sounds without background noise. Further testing using more background noise and using it for training would be beneficial since the goal is an app that functions in daily life that has background noise.

The future step for the project is to optimize the FFNN and CNN further for higher accuracy and test the minimum amount of training samples that can be used without a significant decrease in accuracy. The FFNN and CNN are the neural networks that will be optimized in the future because the RNN was less accurate for all noise backgrounds. To optimize the networks, dropout layers will be added, and different compilers and loss functions will be tested to see what works best for each of the two neural networks. Ultimately, our goal is to develop an application for either a smartphone or smartwatch that will alert the user of different common sounds and allow them to upload their own custom sounds and share them with other users.

## MATERIALS AND METHODS

The audio backgrounds and bell sounds were all taken from online datasets, and the guitar sounds and knocking sounds were recorded using a Blue Snowball microphone (16). The talking sounds were from the Mozilla Common Voice dataset (17).

To test the five human participants, a Python program was developed using the playsound library and Tkinter Python interface (18,19). For the first part of the program, each participant took an audio reaction time test five times, and the average reaction time for each participant was determined. The participant next listened to examples of each of the different sound categories. Then, the participants were asked to categorize five mixed sounds for each of the four background types. The sounds and the background audio were both played at the same time to combine them. The average reaction time of the participant was subtracted from each time to categorize the sound. The accuracy of the participant was the percentage of the chosen category that was the same as the correct category. To find the accuracy for each noise background, the same process was repeated with the subset of sounds for that respective background.

Three types of neural networks were created using the Keras library: FFNN, RNN, and CNN (20). Since the neural networks use numerical data as inputs, the audio of the sounds with the different background types was converted to mel spectrograms using librosa and Matplotlib (21,22). Each of the images of the mel spectrograms was then converted into two-dimensional arrays of the RGB values of the pixels.

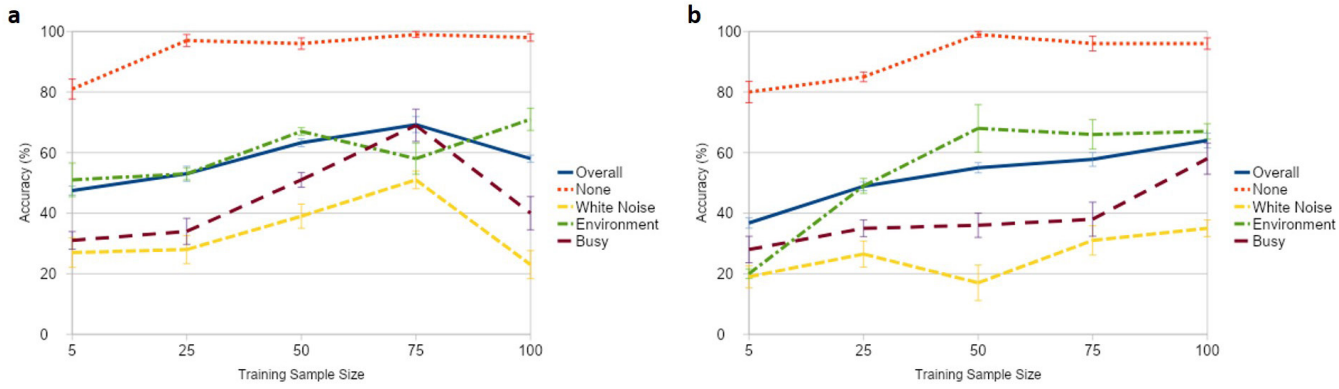


**Figure 4: Accuracy of FFNN and CNN with different training sample sizes for each background type. A)** Average percent accuracy of FFNN (n=5) with training sample sizes of 5, 25, 50, 75, and 100 for overall performance and for each background type. Error bars represent standard deviation. Analysis with two-way ANOVA of training sample size, background noise and accuracy revealed statistical significance with $p < 0.001$. Results of Tukey test comparisons shown in Table 3. **B)** Accuracy of CNN with different training sample sizes for each background type. Average percent accuracy of CNN (n=5) with training sample sizes of 5, 25, 50, 75, and 100 for overall performance and each background type. Error bars represent standard deviation. Analysis with two-way ANOVA of training sample size, background noise and accuracy revealed statistical significance with $p < 0.001$. Results of Tukey test comparisons shown in Table 4.

https://doi.org/10.59720/23-214

| | | 5 | 25 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|
| FFNN | 5 | x | p > 0.05 | p < 0.05 | p < 0.05 | p < 0.05 |
| | 25 | No | x | p < 0.05 | p < 0.05 | p > 0.05 |
| | 50 | p < 0.05 | p < 0.05 | x | p > 0.05 | p > 0.05 |
| | 75 | p < 0.05 | p < 0.05 | p > 0.05 | x | p < 0.05 |
| | 100 | p < 0.05 | p > 0.05 | p > 0.05 | p < 0.05 | x |
| CNN | 5 | x | p < 0.05 | p < 0.05 | p < 0.05 | p < 0.05 |
| | 25 | p < 0.05 | x | p > 0.05 | p < 0.05 | p < 0.05 |
| | 50 | p < 0.05 | p > 0.05 | x | p > 0.05 | p < 0.05 |
| | 75 | p < 0.05 | p < 0.05 | p > 0.05 | x | p > 0.05 |
| | 100 | p < 0.05 | p < 0.05 | p < 0.05 | p > 0.05 | x |

**Table 3: Statistically significant pairs of sample sizes for FFNN and CNN categorization accuracy.** Table with the pairs of sample sizes and whether the difference in their categorization accuracy by the FFNN or CNN is statistically significant.

Each neural network was trained on 100 sounds from each of the four noise categories with no background noise across five epochs, and compiled using the TensorFlow Adam compiler (20). The FFNN used two reLU activation layers and one softmax activation layer. The RNN had two LSTM recurrent layers that used the reLU activation function, one dense layer that used the reLU activation layer, and one dense layer that used the softmax function. The CNN had a convolution layer, a max pooling layer, a flatten layer, and a softmax function. The model of each neural network was exported after compiling.

To test each neural network, each model predicted the category of 100 sounds for each of the four background types, performing 400 predictions. Similarly to the participants, the time, correct category, and predicted category from each neural network were recorded. The average accuracy of the neural networks was found by comparing the predicted category and the correct category. The times and accuracy of the neural networks and participants were then compared to find which one was the fastest and most accurate. To test the statistical significance of the effect of background noise and neural network type on accuracy and speed, a two-way ANOVA test was used. To test the significance of the effect of background noise on the speed of the human participants, a one-way ANOVA followed by the Tukey method was used to find significant differences between pairs.

**Received:** August 23, 2023
**Accepted:** February 23, 2023
**Published:** August 23, 2024

## REFERENCES

1. Temming, Maria. “A Smartwatch App Alerts Users With Hearing Loss to Nearby Sounds.” Science News, 6 Nov. 2020, www.sciencenews.org/article/smart-watch-app-hearing-loss-sounds. Accessed 5 Oct. 2022.
2. Voigt, Ingo et al. “A deep neural network using audio files for detection of aortic stenosis.” Clinical cardiology vol. 45,6 (2022): 657-663. https://doi.org/10.1002/clc.23826.
3. IBM. “What Are Neural Networks? | IBM.” Www.ibm.com, 2021, www.ibm.com/topics/neural-networks. Accessed 5 Oct. 2022.
4. Brownlee, Jason. “How Much Training Data Is Required for Machine Learning?” *Machine Learning Mastery*, 23 July 2017, machinelearningmastery.com/much-training-data-required-machine-learning/. Accessed 20 Mar. 2023.
5. McCullum, Nick. “Deep Learning Neural Networks Explained in Plain English.” FreeCodeCamp.org, 28 June 2020, www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/. Accessed 10 Dec. 2022.
6. “Feed Forward Neural Network.” DeepAI, 17 May 2019, deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network. Accessed 13 Dec. 2022.
7. IBM Cloud Education. “What Are Recurrent Neural Networks?” 14 Sept. 2020, www.ibm.com/cloud/learn/recurrent-neural-networks. Accessed 20 Dec. 2022.
8. Mishra, Mayank. “Convolutional Neural Networks, Explained - Towards Data Science.” Medium, 15 Dec. 2021, towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939. Accessed 11 Jan. 2023.
9. Krishna, B. Vamsi, and T. P. Anithaashri. “Novel Predictive Analysis for Leaf Diseases Using Convolutional Neural Networks Comparing over Artificial Neural Networks.” *4th International Conference on Material Science and Applications*, 2023, https://doi.org/10.1063/5.0172887.
10. Prudhvi Kumar, B., and Anithaashri TP. “Novel Diagnostic System for COVID-19 Pneumonia Using Forward Propagation of Convolutional Neural Network Comparing with Artificial Neural Network.” *ECS Transactions*, vol. 107, no. 1, 24 Apr. 2022, pp. 13797–13814, https://doi.org/10.1149/10701.13797ecst.
11. mlearnere. “Learning from Audio: Spectrograms.” Medium, 14 Apr. 2021, towardsdatascience.com/learning-from-audio-spectrograms-37df29dba98c. Accessed 13 Oct. 2022.
12. Tha Secret. “10 Min White Noise for Stress.” YouTube, 10 Nov. 2015, www.youtube.com/watch?v=NZs-WK3DYpQ. Accessed 4 Jan. 2023.
13. Audio Library - Free Sound Effects. “Warm Afternoon Outdoors - Sound Effect.” *YouTube*, 29 Jan. 2016, www.youtube.com/watch?v=13_BEz2f58A. Accessed 4 Jan. 2023.
14. Audio Library - Free Sound Effects. “Coffee Shop - Sound Effect.” *YouTube*, 29 Jan. 2016, www.youtube.com/watch?v=Plaz8uTGt2w. Accessed 4 Jan. 2023.
15. Yadav, Harsh. “Dropout in Neural Networks.” Medium, 5 July 2022, towardsdatascience.com/dropout-in-neural-networks-47a162d621d9. Accessed 11 Dec. 2022.
16. *Freesound*, freesound.org/. Accessed 4 Jan. 2023.
17. “Common Voice by Mozilla.” Commonvoice.mozilla.org, commonvoice.mozilla.org/en. Accessed 4 Jan. 2023.
18. github.com/bmicc27/neuralNetworkResearch Accessed 27 Oct. 2023.
19. Marks, Taylor, Playsound:1.3.0, 23 Jul. 2021, pypi.org/project/playsound/ Accessed 16 Mar. 2023
20. Tensorflow:2.14.0, 26 Sep. 2023, www.tensorflow.org/. Accessed 11 Oct. 2022.
21. McFee, B., et al. Librosa/librosa: 0.10.1. 0.10.1, Zenodo, 16 Aug. 2023, https://doi.org/10.5281/zenodo.8252662.
22. Matplotlib:3.8.0, 13 Sep. 2023, matplotlib.org/stable/. Accessed 14 Mar. 2023

https://doi.org/10.59720/23-214