Identification of Patients with Congenital CMV (cCMV) Infection through a Deep Neural Network

Alexandra Medoro MD

Background:

Congenital cytomegalovirus (cCMV) infection is the most common congenital viral infection affecting ~0.4-1% of all live birth each year in the United States and over a million newborns globally.1 Due to the prevalence of seropositivity to cytomegalovirus in the general population, it is generally agreed that the diagnosis of congenital infection needs to be made before three weeks of age2. A primary sequala of interest in cCMV infection is sensorineural hearing loss. While this hearing loss may be present at birth, it can also present later in childhood. This late presentation of symptomatology combined with the time-sensitive nature of cCMV diagnosis leads to a population of patients with sensorineural hearing loss of unclear etiology and possibly from cCMV infection. The ability to classify a patient as either cCMV infected or not was the motivation for this project. Our objective was to construct a deep neural network (DNN) to classify a known population of patients.

Methods:

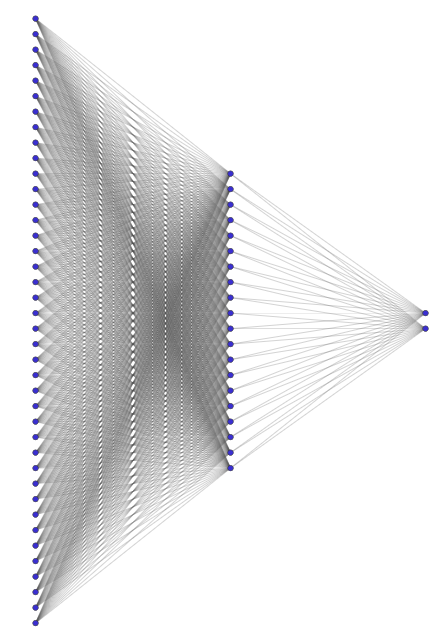
RNA-seq data for 49 subjects was obtained. Of these subjects, 39 had cCMV infection confirmed by positive DNA PCR result within 21 days of age or a positive dried blood spot test. The remaining 10 subjects were used as controls. Statistical analyses were performed using GraphPad Prism version 8.0.0 for Windows, GraphPad Software, San Diego, California USA, [www.graphpad.com](http://www.graphpad.com). The dataset was divided into a training set and a test set using the R package *caTools* with a sample split of 2/3 of samples into the training set and 1/3 of samples into the test set. The control samples were used in both samples (FIGURE 1). Demographic data were recorded on both the cCMV infected subjects and the control subjects and comparisons were done between the subject and control groups (TABLE 1) and the training and test groups (TABLE 2).

Figure 1. Patient schema of the experimental design dividing the subjects into a training and a test set.

Utilizing the R package *DESeq2*, differential expression analysis was performed on the training data set3. Significantly differentially expressed genes were designated as those for whom the Benjamini-Hochberg adjusted p-value was less than 0.1 and the absolute value log fold change > 1.5. From the original 58,288 genes in the data set, 831 genes met these criteria and thus comprised the input features for the deep learning model. This gene signature was mapped across both the training and test samples following a variance stabilizing transformation with normalization to the control subjects.

A DNN model was constructed utilizing TensorFlow. The first version of the model utilized only an input layer and an output layer with no hidden layer. The input layer consisted of the expression of the 831 significantly differentially expressed genes. The output layer represented a classification of the subject as either “cCMV infected” or “control” using the sigmoid activation function. The learning rate was chosen as 0.05 and 50 epochs were run with a batch size of 10 (FIGURE 5). These parameters were chosen through a process of serial selection. To improve the accuracy of the model, a hidden layer of 20 neurons utilizing the relu activation function was added with the number of neurons chosen through serial selection (FIGURE 6) using the same parameters as the previous model. The final model was created through the addition of a drop out rate in both the input layer and the hidden layer (FIGURE 7). The schematic of the final model is shown below in FIGURE 2.

Figure 2. Schematic of DNN model with an input layer of 831 features (differentially expressed genes), a hidden layer of 20 neurons and an output layer of two neurons.



Results and Discussion:

The demographic characteristics of the subjects with congenital CMV infection and the control patients are shown in TABLE 1. There was no significant different in the subject and control group with respect to sex or race distribution. However, there were significant differences in gestational age, birth weight, and age of sample collection. Overall, this is not unexpected. Congenital CMV infection is associated with low birth weight due to intrauterine growth restriction as well as premature birth, thus we were not surprised that our subjects and controls differed in these areas. The difference in age was a limitation of study design given the difficulty in obtaining blood samples from well infants due to the global pandemic which occurred during the study period. The demographic characteristics of the training and test group are shown in TABLE 2. There was no significant difference in any demographic characteristic between these two groups.

*Table 1. Demographic characteristics of subjects with congenital CMV infection and control patients.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Congenital CMV** | **Control** | **p value** |
| n | 39 | 10 |  |
| Gestational Age (weeks), median [IQR]  Range | 37 [36-39]  28-40 | 39 [37.5-41]  36-41 | 0.01 |
| Birth Weight (grams), median [IQR]  Range | 2325 [1776-2882]  990-3544 | 3522 [2563-4058]  1511-4340 | 0.0008 |
| Age (days), median [IQR]  Range | 11 [7-19]  2-234 | 57.5 [54.5-63]  42-68 | <0.0001 |
| Sex (n, %)  Female  Male | 13, 33%  26, 67% | 7, 70%  3, 30% | 0.08 |
| Race (n, %)  White  Black or African American  Other | 23, 56%  10, 24%  8, 20% | 7, 70%  2, 20%  1, 10% | 0.62 |

Table 2. Demographic characteristics of subjects divided into the training and test group.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Training** | **Test** | **p value** |
| n | 36 | 23 |  |
| Gestational Age (weeks), median [IQR]  Range | 37.5[36-39]  28-41 | 38[36-39]  28-41 | 0.9719 |
| Birth Weight (grams), median [IQR]  Range | 2551[2100-3418]  990-4340 | 2580[1511-3487]  1170-4340 | 0.9591 |
| Age (days), median [IQR]  Range | 19.5[10-55]  2-234 | 16[8-58]  4-74 | 0.9969 |
| Sex (n, %)  Female  Male | 17, 47%  19, 53% | 10, 43%  13, 57% | 0.7959 |
| Race (n, %)  White  Black or African American  Other | 22, 61%  9, 25%  5, 14% | 13, 57%  5, 22%  5, 22% | 0.7324 |

PCA plots of both the training and test data are shown in FIGURE 3 displaying the first and second principal component for each data set. Control subjects are shown in pink and the disease subjects are shown in blue and green. The training and the test PCA plots shown similar dispersal patterns. The variance stabilization transformed data were plotted in a heat map (FIGURE 4) after selecting for the significantly differentially expressed genes. The first panel of the figure are the control subjects. The second panel of the figure are the patients in the training set and the third panel of the figure are the patients in the test set. Genes which are overexpressed are shown in red and genes which are under expressed are shown in blue. The expression patterns of the training and test sets are visually similar and visually different from those of the control subjects.

Figure 3. PCA plots of training data (left) and test data (right). Control patients are shown in pink while patients with cCMV are shown in blue and green.

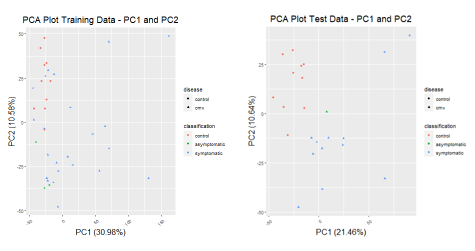
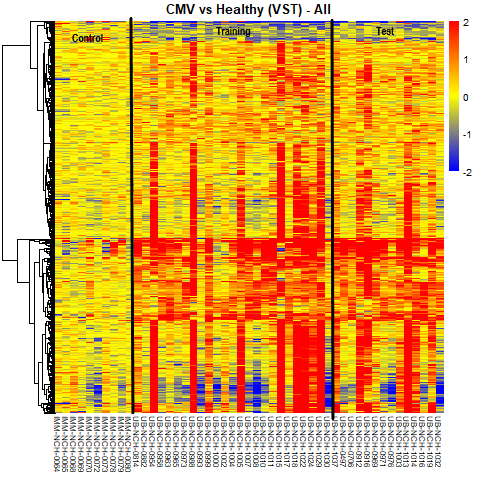
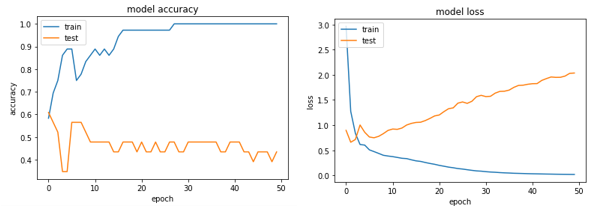


Figure 4. Heat map of significantly differentially expressed genes defined in the training set as a Benjamini-Hochberg adjusted p-value of < 0.1 and an absolute value log fold change > 1.5. Differential expression analysis was performed using DESeq2 R package and a variance stabilizing transformation.



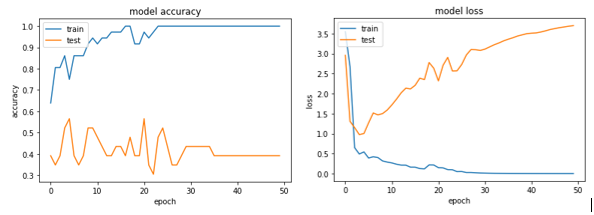
The expression profiles shown in FIGURE 4 served as the features of the input layer for the deep neural network. The final model was constructed in a stepwise method. The first version of the model (FIGURE 5) consisted of only an input layer and an output layer. Model accuracy for the training set approached 100%; however, model accuracy for the test set decreased across the epochs. In addition, while the loss function decreased across epochs for the training data, it increased for the test data. Thus, additional complexity was needed.

Figure 5. Model 1. DNN model accuracy and loss plotted as a function of the epoch. This model consisted of an input layer and an output layer using the sigmoid activation function. The learning rate was 0.05. Fifty epochs were run with a batch size of 10. The accuracy of this model on the test set data was 0.4348.



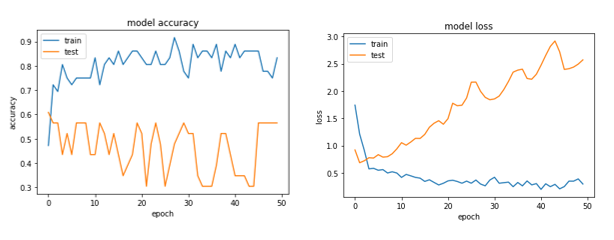
The second version of the model (FIGURE 6) added a hidden layer of 20 neurons which keeping the rest of the model parameters the same. This didn’t immediately have an effect on the model performance accuracy althought the loss function seemed to increase at a steeper rate.

Figure 6. Model 2. DNN model accuracy and loss plotted as a function of the epoch. This model consisted of an input layer, a hidden layer of 20 neuron using the relu activation function, and an output layer of two neurons using the sigmoid activation function. The learning rate was 0.05. Fifty epochs were run with a batch size of 10. The accuracy of this model on the test set data was 0.3913.



The final version of the model incorporated a drop out rate in both the input layer and the hidden layer in an attempt to compensate for the overfitting shown in the previous two versions of the model. This was marginally successful as the final version of the model now has an accuracy of 56% for the test data. While the loss function still increases as the epochs progress, indicating residual overfitting, it increases at a shallower rate than previous model versions.

Figure 7. Final Model. DNN model accuracy and loss plotted as a function of the epoch. This model consisted of an input layer with a dropout rate of 0.2, a hidden layer of 20 neurons with a dropout rate of 0.5 using the relu activation function, and an output layer of two neurons using the sigmoid activation function. The learning rate was 0.05. Fifty epochs were run with a batch size of 10. The accuracy of this model on the test set data was 0.5652.



Conclusions, Limitations, and Future Directions:

This model demonstrates a proof of concept of the idea of disease classification through the use of a deep neural network. Though ideal accuracy has not yet been achieved, the concept is likely valid due to the difference in gene expression data pattern between control subjects and disease subjects demonstrated in the input features. It is possible that a more complex network that takes into account relationships between the individual features may better describe these data and is certainly an area for future development.

A limitation of these data is the difference in age of the disease and control subjects. The maturation of the neonatal immune system is still not well described and so it is possible that there may be a significant signal difference due to age of subject alone. It may be possible in future models to train based on age of subject as well in order to account for this confounding factor.

Attachments:

Juptyr Notebook with Model Code

Data Files

* test\_features.csv : differentially expressed genes for test data set
* test\_targets\_scored.csv: subject identification for test data set
* train\_features.csv: differentially expressed genes for training data set
* train\_features\_scored.csv: subject identification for training data set

References

1. Kimberlin DW, Lin CY, Sanchez PJ, et al. Effect of ganciclovir therapy on hearing in symptomatic congenital cytomegalovirus disease involving the central nervous system: a randomized, controlled trial. *J Pediatr.* 2003;143(1):16-25.

2. American Academy of Pediatrics. Cytomegalovirus Infection. In: Kimberlin DW BM, Jackson MA, Long SS, ed. *Red Book: 2018 Report of the Committee on Infectious Diseases.* American Academy of Pediatrics; 2018:310-317.

3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.