Neural Network Classification of Depressive Disorder Based on Prior Stressful Life Events

Kira Breeden

March 13, 2021

kmbreeden21@gmail.com

Abstract

Depressive disorders are becoming increasingly prevalent in our world today and identification of possible symptoms and indicators can be life-saving. Given the occasional lack of consistency in physicians' diagnoses of depression, we must look to other methods to help identify possibly dangerous depressive tendencies. This paper tackles this problem using artificial neural networks to analyze BRFSS survey data collected but the CDC in order to classify depressive disorder in respondents. This paper investigates the use of a multilayer perceptron neural network architecture for depressive disorder classification while varying the input features. We preprocessed and analyzed the data from the BRFSS dataset and narrowed down possible depression indicators using correlation analyses and prior literature that identifies "stressful life events" that may cause depression (Harvard Health Publishing 2019). Experimental results show that features such as income, poor mental health feelings, and preexisting health conditions may have larger predictive power when identifying the existence of depressive disorders. These results indicate that our architecture using features tracking semi-specific life events is most accurate when classifying instances of these disorders.

Keywords: Depressive Disorders, Multilayer Perceptron, Neural Networks, Telephone Surveys, Stressful Life Events.

Introduction

Instances of depressive disorders have been rising in the United States over time as outlined by Weinberger et al (2017). We find this rise in depressive disorder particularly noticeable in the youngest and oldest age groups (Weinberger et al. 2017). Artificial neural networks have been an increasingly popular method of choice when identifying cases and symptoms of depressive disorder (Lin 2020) and can give us a unique look into the relationships between different aspects of life and how they affect our mental health. Given the many variables that may cause depression — including but not limited to genetics, stress, trauma, and medical problems — the introduction of nonlinearity in neural networks provides promising predictive power (Harvard Health Publishing 2019). Not only are the possible causes of depressive disorder numerous but the process of diagnosis for these disorders contains many nuances and disagreements among professionals (Mitchell, Vaze, & Sanjay 2009). For this reason, the use of neural networks to provide insights into critical features that may predict depression is increasingly important and valuable in our society. For the problem at hand, we will be tuning a multilayer perceptron architecture to determine the most important features for identifying depressive disorder based on previous “stressful life events” (Harvard Health Publishing 2019).

Literature Review

There are two common ways in which artificial intelligence has been used to identify and predict depressive disorders in the world’s population. The first is in the use of neural networks to identify valuable features from psychological and medical datasets (Suhara et al. 2017). This method provides a strong backbone for identification of depressive disorders (provided the dataset is robust). In this case, we require a reliable labeling method for the disorders, which means we require that the researchers who collected the data are consistent in their diagnoses and coding methods. This method is displayed by Suhara et al (2017) in their use of an app-based tracking of mood among study participants. This was a method of self-reporting where users would input their mood throughout the day (at any time they wished) and the researchers used convolutional neural networks (CNNs) to predict levels of depressive feelings based on mood analysis throughout the day (2017). In this case, the infrastructure of a self-reporting app-based data collection is very clean and does not involve the issue of mistakes in diagnoses by professionals. The use of CNNs in this case was largely made possible by the ability to input the data as a time-series of mood throughout the day.

The second common method for depressive disorder classification is the use of Natural Language Processing (NLP) to analyze text sources to determine sentiment that may be linked with depressive disorders and depressive feelings. This method is most often used in tandem with people's interactions on social media platforms (like twitter, Reddit, and Facebook). Some notable works in this are are those by Gkotsis et al. (2017) and Lang He and Cui Cao (2018) where they use combinations of CNNs and RNNs in natural language processing to determine sentiment that may be linked with depression. Additionally, the work by Tung Tran and Kavuluru Ramakanth (2017) showcases the use of NLP to analyze medical notes made by professionals referring to patients with depressive disorders once again using CNN and RNN methods.

Overall, the history of classification problems in regard to depressive disorders use CNNs and RNNs most frequently when looking at data that may be interpreted as time-series data or recurrent predictions (like analyzing doctor's notes word by word). For this project, we do not have data that can be interpreted as time series data so we will be using a simple multilayer perceptron with careful consideration for the best predictive features. In particular, we want to determine whether features that are classified as “stressful life events” actually provide predictive power for simple multilayer perceptron models. This could be extremely useful especially if people are working with less extensive data. For example, if there was a school that tracks the home situations and life events of their students, would a simple multilayer perceptron be able to provide enough predictive power to successfully provide more support to students in need?

Data

The data used for this study was collected by the Center for Disease Control and Prevention’s (CDC’s) Population Health Surveillance Branch throughout the years of 2016-2019 as a part of the Behavioral Risk Factor Surveillance System (Center for Disease Control and Prevention 2020). The BRFSS is “the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services” (Center for Disease Control and Prevention 2020). The dataset contains all survey responses from 2016-2019 for all respondents to a phone survey performed across the United States. Each response was coded numerically into a category described by the BRFSS codebook (Behavioral Risk Factor Surveillance System 2020). This dataset contains all the aggregate responses from land line and cell phone surveys from 49 states, the District of Columbia, Guam, and Puerto Rico. The state of New Jersey did not collect enough data to be included in the dataset. Each of the states and territories that participate in the survey process submit all survey responses each month to the CDC to be coded into the working dataset (Center for Disease Control and Prevention and Behavioral Risk Factory Surveillance System 2019).

The age requirements for the BRFSS survey response is 18 years old and above, therefore data for depression and other health conditions in minors in not included in this dataset. “Factors assessed by the BRFSS in 2019 included health status, healthy days/health-related quality of life, health care access, exercise, inadequate sleep, chronic health conditions, oral health, tobacco use, e-cigarettes, alcohol consumption, immunization, falls, seat belt use, drinking and driving, breast and cervical cancer screening, prostate cancer screening, colorectal cancer screening, and HIV/AIDS knowledge” (Center for Disease Control and Prevention and Behavioral Risk Factory Surveillance System 2019).

The main dataset was downloaded from the BRFSS website and converted from the SAS Transport file format to comma separated values format.

Methods

We began by determining which feature from the BRFSS dataset we wanted to use as our label in our classification model. The survey tracked whether respondents had a history of depressive disorder (diagnosed by a professional) and the respondents answers were threefold: “yes”, “no”, or “I’m not sure.” This three-part classification label is what we decided to use for our label in our neural networks keeping in mind that this is not predicting *current* cases of depressive disorder but merely the existence of depressive disorders somewhere in the person’s history.

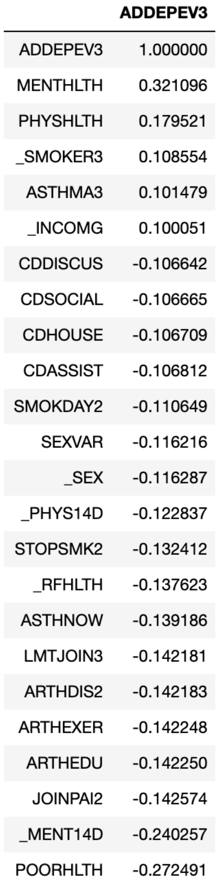
After determining the label feature for our problem, we began the preprocessing and cleaning of the original BRFSS dataset. Conveniently, most of the data had already been coded as numerical values corresponding to the answers in the BRFSS codebook (Behavioral Risk Factor Surveillance System 2020). We began by replacing all the values of "." (which was used to indicate when the question was not asked of the respondent in the original survey) with the dummy value of -1000. After this replacement, we confirmed that all the values in each column were numerical by mapping all string inputs to numerical values. Next we dropped all the cases where the researchers did not ask the question about depression to the respondent (seeing as this response variable is acting as our label for our classification).

Once the necessary preprocessing was complete and our label feature was determined, we moved on to performing some EDA on the cleaned dataset in order to gain more powerful insights into the data. We closely looked at the correlations between different features and the response variable using a correlation heat map. In this we determined that the strongest correlated features are as shown in Figure 1.

The BRFSS dataset originally contained over 300 features, most of which were not related to the problem at hand at all. For this reason, we stripped down the dataset to one containing features regarding health, depressive disorders, demographics, and other features that could be considered "stressful life events" (Harvard Health Publishing 2019). In this process, we also ensured that the features that were highly correlated to our depression label were included in the main data frame we would use moving forward.

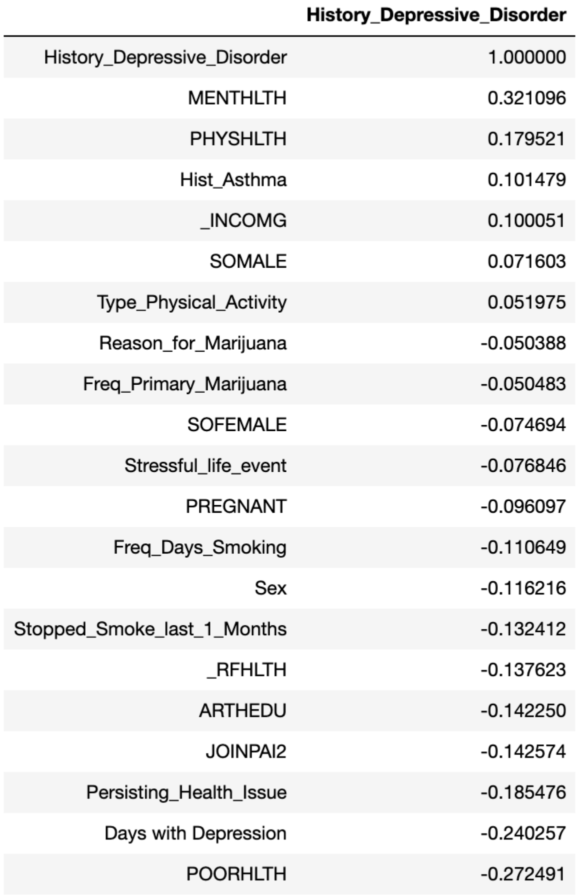
We then continued into some feature engineering for the database on hand. In order to test whether simply having information about the existence of a "stressful life event" is useful in predicting depression, we engineered two binary features. One was simply a binary feature marking the existence of a previous stressful life event. This counted as having experienced one or more of the following: sexual abuse, parental violence, or parental divorce. The next feature was just the presence of a persisting health issue. This was determined by the existence of one or more of the following: arthritis, joint pain, asthma, and general poor physical health more than half the time.

These two features were titled “Stressful_life_event” and “Persisting_Health_Issue” respectively. For each, we used 1 and 0 as the values of the feature where 1 indicated the existence of one or more stressors or health issues, and 0 indicated no stressors or health issues. The updated top correlations with the inclusion of the engineered features are shown in Figure 2.



	ADDEPEV3
ADDEPEV3	1.000000
MENTHLTH	0.321096
PHYSHLTH	0.179521
_SMOKER3	0.108554
ASTHMA3	0.101479
_INCOMG	0.100051
CDDISCUS	-0.106642
CDSOCIAL	-0.106665
CDHOUSE	-0.106709
CDASSIST	-0.106812
SMOKDAY2	-0.110649
SEXVAR	-0.116216
_SEX	-0.116287
_PHYS14D	-0.122837
STOPSMK2	-0.132412
_RFHLTH	-0.137623
ASTHNOW	-0.139186
LMTJOIN3	-0.142181
ARTHDIS2	-0.142183
ARTHEXER	-0.142248
ARTHEDU	-0.142250
JOINPAI2	-0.142574
_MENT14D	-0.240257
POORHLTH	-0.272491

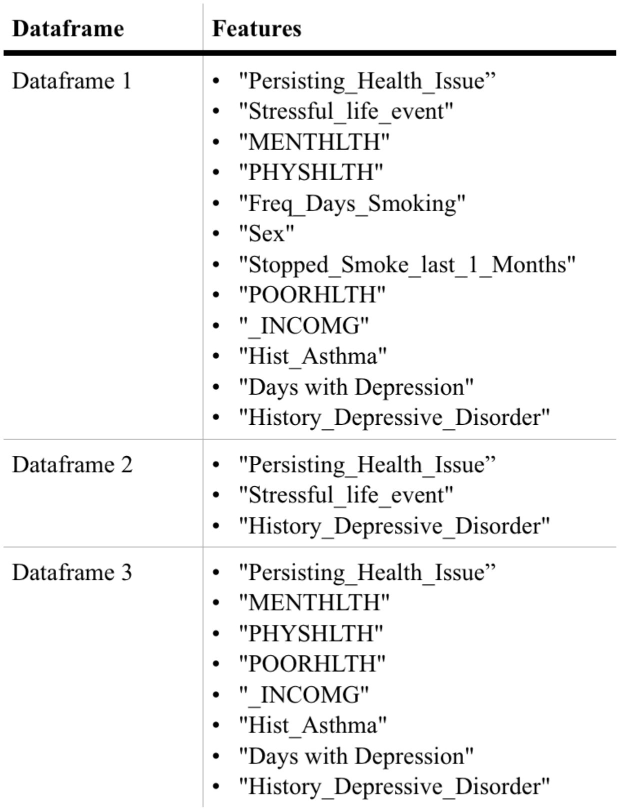
Figure 1. Table of features strongly correlated with depressive disorder label (ADDEPEV3) from original datset.



	History_Depressive_Disorder
History_Depressive_Disorder	1.000000
MENTHLTH	0.321096
PHYSHLTH	0.179521
Hist_Asthma	0.101479
_INCOMG	0.100051
SOMALE	0.071603
Type_Physical_Activity	0.051975
Reason_for_Marijuana	-0.050388
Freq_Primary_Marijuana	-0.050483
SOFEMALE	-0.074694
Stressful_life_event	-0.076846
PREGNANT	-0.096097
Freq_Days_Smoking	-0.110649
Sex	-0.116216
Stopped_Smoke_last_1_Months	-0.132412
_RFHLTH	-0.137623
ARTHEDU	-0.142250
JOINPAI2	-0.142574
Persisting_Health_Issue	-0.185476
Days with Depression	-0.240257
POORHLTH	-0.272491

Figure 2. Table of features strongly correlated with depressive disorder label with the inclusion of the engineered features from the narrowed down datset.

Next, we created three separate databases with different feature combinations to test and see which combinations of features provided the most predictive power. The breakdown of the features in each dataset is shown in Figure 3.

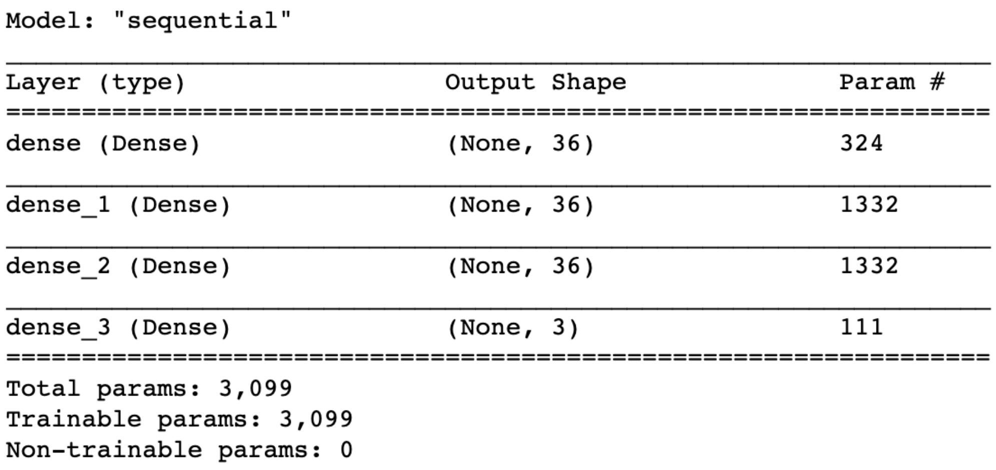


Dataframe	Features
Dataframe 1	• "Persisting_Health_Issue” • "Stressful_life_event" • "MENTHLTH" • "PHYSHLTH" • "Freq_Days_Smoking" • "Sex" • "Stopped_Smoke_last_1_Months" • "POORHLTH" • "_INCOMG" • "Hist_Asthma" • "Days with Depression" • "History_Depressive_Disorder"
Dataframe 2	• "Persisting_Health_Issue” • "Stressful_life_event" • "History_Depressive_Disorder"
Dataframe 3	• "Persisting_Health_Issue” • "MENTHLTH" • "PHYSHLTH" • "POORHLTH" • "_INCOMG" • "Hist_Asthma" • "Days with Depression" • "History_Depressive_Disorder"

Figure 3. Table of input features associated with each data frame.

We then split our dataset into our training, validation, and test sets and one hot encoded all the values of each feature. This provided us with the necessary preparation for feeding our data into our neural network architecture for each data frame.

We trained our model three separate times using the three different selections of input features. The only variation in our model between trainings was a difference in the input size (to account for a different number of features in each data frame). The use of a sigmoid activation function for each layer of the network was chosen in order to introduce a nonlinearity to the predictive power of our networks. The network architecture that were used is shown in Figure 4.

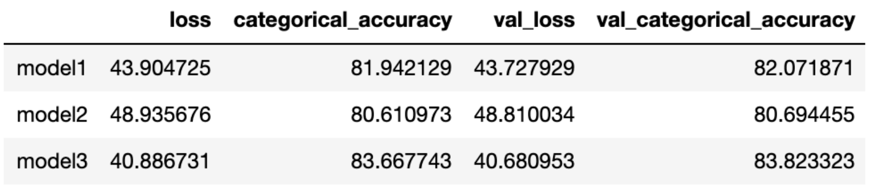


```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense (Dense)                (None, 36)                324       
_________________________________________________________________
dense_1 (Dense)              (None, 36)                1332      
_________________________________________________________________
dense_2 (Dense)              (None, 36)                1332      
_________________________________________________________________
dense_3 (Dense)              (None, 3)                 111       
=================================================================
Total params: 3,099
Trainable params: 3,099
Non-trainable params: 0
```

Figure 4. Multilayer Perception Architecture

Results

After testing our three different feature combinations, we evaluated the categorical cross-entropy loss and the categorical accuracy for each training. Each training of our model is marked in Figure 5 as “model 1,” “model 2,” or “model 3” in order to indicate the necessary change in input size of the existing model in Figure 4. Figure 5 below shows the final loss and accuracy for each training for both the training and validation sets.



	loss	categorical_accuracy	val_loss	val_categorical_accuracy
model1	43.904725	81.942129	43.727929	82.071871
model2	48.935676	80.610973	48.810034	80.694455
model3	40.886731	83.667743	40.680953	83.823323

Figure 5. Final Model Accuracy and Loss Measures

Additionally, we tracked the relationship between the test set and the validation set performance over each epoch of the models. The resulting graphs are shown below in Figures 6 - 11.

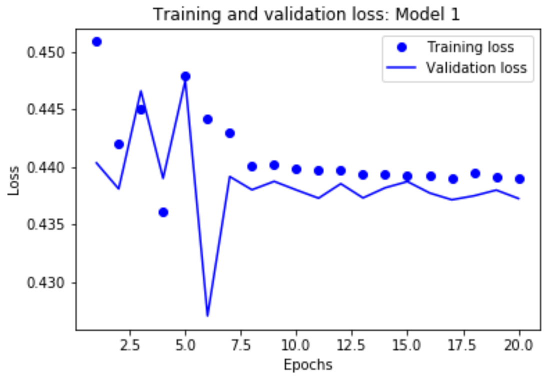


Figure 6. Loss over time for Model 1

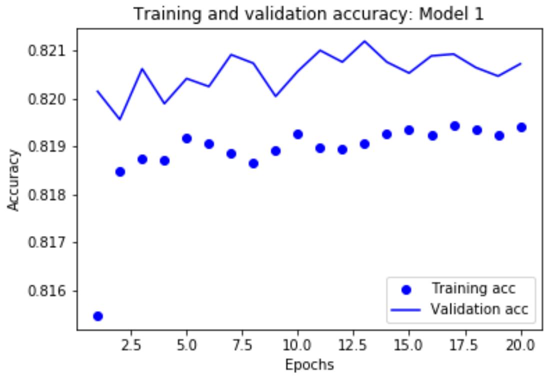


Figure 7. Accuracy over time for Model 1

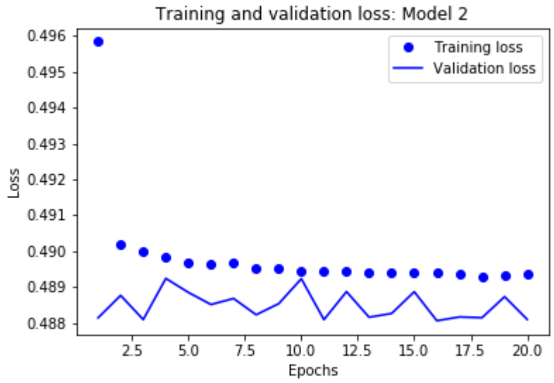


Figure 8. Loss over time for Model 2

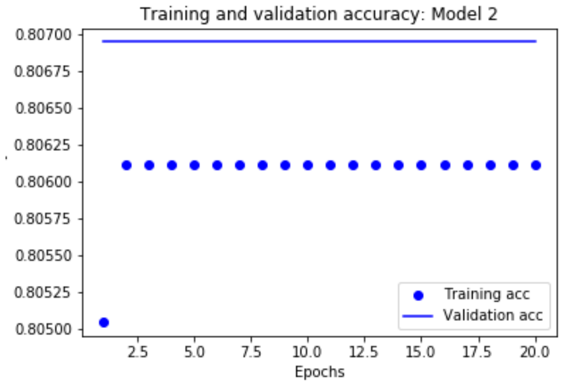


Figure 9. Accuracy over time for Model 2

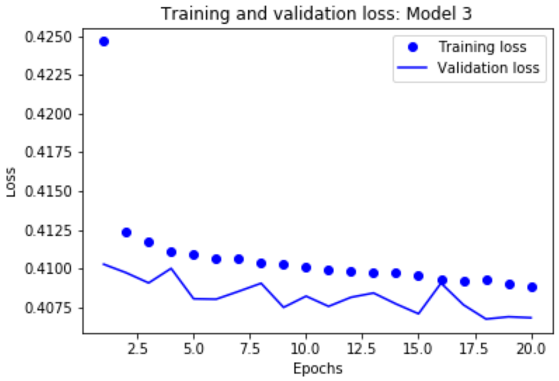


Figure 10. Loss over time for Model 3

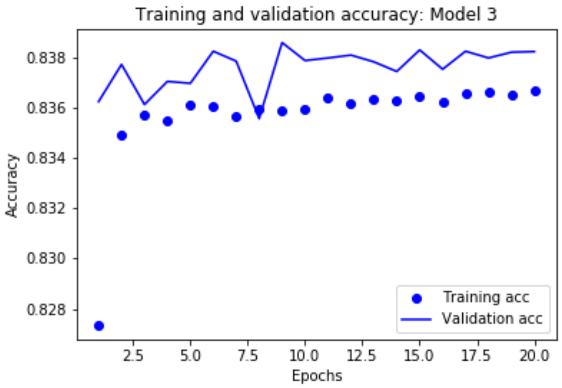


Figure 11. Accuracy over time for Model 3

Analysis and Interpretation

In regard to the interpretation of our results, we first need to consider the difference between the performance of the model on the training set versus the validation set. As we can see from the above results section, our models achieve an accuracy that is very similar between training and validation sets (cf. Figure 5). This means that we are neither overfitting nor underfitting our data which is a pretty great achievement. This is most likely an effect of the consistency of the survey questions and coding responses. Having our data be pre-coded by the researchers developing the survey provided data with little complexity.

The next thing we need to consider is how the difference in features changed the accuracy of our algorithm. Starting with model 1: we used the same multilayer perceptron architecture and we included all of the features associated with higher correlation to a history of depressive disorders (cf. Figure 3). Often, when choosing features to include in a neural network architecture, we need to keep in mind Occam's Razor: the simplest of competing theories is often preferred. Too many features can often create too much complexity for the algorithm to sift through and can make it much harder for the layers to glean predictive information. This turns out to be the case with our network as well. We can see that model 1 performs decently with a validation set accuracy score of 82.07%.

With model 2, we stripped down all the features to only the two that we engineered. Recall that these features were binary features indicating the existence of either a persisting health issue or a stressful life event. In this case, the intention was to see whether avoiding specifics of *what* health issue or life event had taken place, was enough to provide acceptable predictive accuracy. As we can see from our results, our accuracy did go down to 80.69% from that of model 1 (82.07%) where all of our features were included. This does indicate that providing specific information about mental health and physical health experiences does lead to higher model accuracy.

This lead us to our final model where we tried to find the optimal number of features to avoid confusing the model while providing sufficient information. We landed on model 3 where we used 7 input features all of which have a correlation of ≥ 0.1 or ≤ -0.1 with a history of depressive disorder. With this model we can see that we ended up with a validation set accuracy of around 83% which is higher than both model 1 and 2. We notice that specific information about health issues (such as asthma and joint pain), income, and poor mental health can greatly aid the predictive power of our model.

Conclusions

Overall we can see that the choice of features to feed into our model matters quite a lot in the accuracy of the model. This study shows the importance of the relationship between people's life experiences and their mental health. In particular, this study may be of considerable use for schools in the United States and for depression awareness across the nation. If students' families responded to survey questions about life at home (even without specifics as we saw in model 2) we would be able to predict likelihood of a history of depression in the student with 80.69% accuracy. This could greatly help influence the placement of school counselors, as well as access to healthcare and therapy for children across the nation.

Future Work and Possible Limitations

One limitation of this study is that the survey data may be subject to response bias. This is a very particular dataset from phone surveys across the United States of America and this may not fully represent the general public's responses to such questions or their relationship with depressive disorders. Additionally, when surveys are conducted by a researcher, it is more likely for respondents to lie especially when discussing personal information.

Another possible limitation of this study lies in the label feature itself. In particular, the question asked of the respondents is whether they have had a history of *diagnosed* depressive disorder. This opens the Pandora's box of differing diagnoses and access to healthcare for respondents. In future works, conducting a study where respondents are able to fill out their own responses and where the label feature does not completely rely on professional diagnoses may cast even more light on the relationship between physical and mental health and histories of depression.

Another step to be taken in future works is to simply expand the respondent pool. This could be achieved by creating a phone application where people can respond over time similar to what was done by Suhara et al (2017).

Bibliography

"Behavioral Risk Factor Surveillance System" Center for Disease Control and Prevention. August 31, 2020. https://www.cdc.gov/brfss/index.html

Center for Disease Control and Prevention and Behavioral Risk Factory Surveillance System. "Behavioral Risk Factor Surveillance System. Overview: BRFSS 2019." PDF file. July 26, 2019. https://www.cdc.gov/brfss/annual_data/2019/pdf/overview-2019-508.pdf

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T., Dobson, R., Dutta, R., "Characterisation of mental health conditions in social media using Informed Deep Learning". *Sci Rep* 7, 45141 (2017). https://doi.org/10.1038/srep45141

He, Lang, Cao Cui, "Automated depression analysis using convolutional neural networks from speech" *Journal of Biomedical Informatics* 83 (July 2018): 103-111. https://doi.org/10.1016/j.jbi.2018.05.007

Lin, Chenhao, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. "Sensemood: Depression detection on social media." In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (2020) : 407-411.

"LLCP 2019 Codebook Report" Behavioral Risk Factor Surveillance System. July 31, 2020. https://www.cdc.gov/brfss/annual_data/2019/pdf/codebook19_llcp-v2-508.HTML

Mitchell, Alex J, Amol Vaze, and Sanjay Rao. "Clinical Diagnosis of Depression in Primary Care: a Meta-Analysis." *The Lancet* 374, no. 9690 (August 28, 2009): 609–19. https:/doi.org/10.1016/s0140-6736(09)60879-5.

Suhara, Yoshihiko, Yinzhan Xu, and Alex'Sandy Pentland. "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks." In *Proceedings of* *the 26th International Conference on World Wide Web* (2017): 715-724.

Tran, Tung, and Ramakanth, Kavuluru. "Predicting Mental Conditions Based on 'History of Present Illness' in Psychiatric Notes with Deep Neural Networks." *Journal of Biomedical Informatics* 75 (June 10, 2017). https://doi.org/10.1016/j.jbi.2017.06.010.

"What Causes Depression?" Harvard Health Publishing. June 24, 2019. https://www.health.harvard.edu/mind-and-mood/what-causes-depression.

Weinberger, A. H., M. Gbedemah, A. M. Martinez, D. Nash, S. Galea and R. D. Goodwin. "Trends in Depression Prevalence in the USA from 2005 to 2015: Widening Disparities in Vulnerable Groups." *Psychological Medicine* 48, no 8 (2018) : 1308–1315. doi:10.1017S0033291717002781.