Predictive Modeling of Childhood Internalizing Disorders with KID Study Data

Andrew McCracken
University of Vermont
109 Carrigan Drive, Burlington, VT
05405
andrew.mccracken@uvm.edu

Kevin Motia
University of Vermont
82 University Pl, Burlington, VT
05405
kevin.motia@uvm.edu

Adam Ontiveros
University of Vermont
82 University Pl, Burlington, VT
05405
adam.ontiveros@uvm.edu

ABSTRACT

In the KID Study [3], 84 kids are put into three tasks of varying stress levels while equipped with wearable sensors [4] that record their movement and vocal activity. The purpose of this study is to develop a more efficient means of diagnosing children with disorders like depression, anxiety and ADHD using this data. Although models have been generated in the original study that predict the accuracy of correctly diagnosing a child with an internalizing disorder or ADHD, it can always be improved upon. The goal of the research was to take data from a stress inducing task, create a pipeline including classification models, generalize said model to be able to take data from any task, and acquire better results then those in the KID Study.

Keywords

Machine Learning, Classification Models, Internalizing Disorders, Wearable Sensors

1. INTRODUCTION

It is important for everyone, no matter what age, to have good mental health. Bad mental health, left unchecked, can cause an array of problems, and left unchecked for children and adolescents, can have a negative impact on development [2,5,6], as well as many more problems. Currently, guardians of children are responsible for reporting on the mental health of the child, which can be biased. Children might have a hard time understanding and communicating their more abstract emotions, but the KID study is trying to give children the ability to communicate these emotions though stress responses to varying tasks. These tasks include: playing with a bubble machine for 3 minutes, giving a 3 minute speech where they are interrupted by buzzers twice, and approaching a covered object in a dimly lit room. During the bubbles and approach task, wearable sensors track movement data and during the speech task, audio data is tracked. The counts of each possible combination of diagnoses of the children can be seen in *Figure 1*.

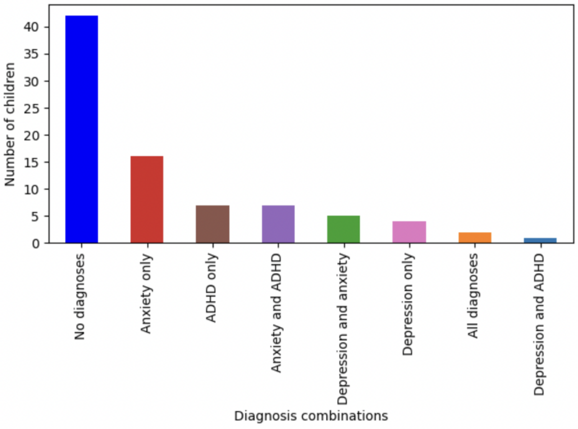
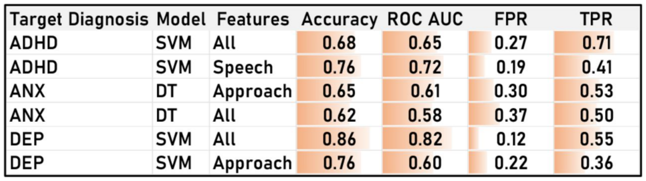


Figure 1: Counts of Diagnosis Combinations

The goal of this study is to generate a pipeline that takes data from the ChAMP app and utilizes machine learning classification methods to predict the diagnosis for children in the future. After the pipelines are created for a particular task, they should be generalized to take data for any of the tasks. The KID Study paper provides a table of the results for the accuracy, ROC AUC, true and false positive rate for the models they used [6]. This table is shown below in Figure 2. The results in Figure 2 are referred to as the benchmark that this project is trying to best.



Target Diagnosis	Model	Features	Accuracy	ROC AUC	FPR	TPR
ADHD	SVM	All	0.68	0.65	0.27	0.71
ADHD	SVM	Speech	0.76	0.72	0.19	0.41
ANX	DT	Approach	0.65	0.61	0.30	0.53
ANX	DT	All	0.62	0.58	0.37	0.50
DEP	SVM	All	0.86	0.82	0.12	0.55
DEP	SVM	Approach	0.76	0.60	0.22	0.36

Figure 2: Table of results from Loftness et al., 2023

2. RELATED WORK

The first piece of this work this project stems off of is the ChAMP app paper [3]. The ChAMP app was developed to deviate away from how child diagnoses historically go, where the guardian of the child speaks on behalf of the child and how the child feels. The ChAMP app is a mobile application that collects movement and audio data during the mood induction tasks mentioned above and also extracts digital biomarkers and digital phenotypes [3]. Continuing, there was work that utilized the ChAMP app for a qualitative study that involved 84 children going through the tasks [4]. This work used a variety of clustering and classification methods to predict the diagnosis of future children, which is what inspired this project. A similar study was performed by Faeda et al [1] where wearable sensors were employed to collect health data during sleep. The purpose of the study was to determine if the collected data could be used to differentiate between children with ADHD, bipolar disorder, and depression. Comorbidity was present in the data collected by Faeda et al. While the study did not influence this project, their models may be worth exploring to inspire future work.

3. METHODS

The data received for this project is the data that was used in the preliminary studies and therefore had already been cleaned [3]. The data set was obtained in CSV format from the M-Sense group at the University of Vermont. There were separate files for the features and target variables. The features were split up into the three tasks which were being examined in the KID study; they are labeled “approach”, “speech” and “bubbles”. After splitting the data into each task, the approach task had 160 features, the Bubbles task had 192 features, and the speech task had 264 features. Even though the data was already cleaned, there was much work to make sense of the features to get usable models.

The target data for the project had columns about the sex and age of the child, if the child was diagnosed with depression, anxiety or ADHD, how many symptoms of each diagnosis they had, and if they have been diagnosed with something, what variant of the disorder they were classified to have. From the target data, three variables were used as targets and to manufacture other target columns for the models: “depdxever”, “anxdxever”, and “adhddxever”. These target variables indicated if the child had depression, anxiety, or ADHD. These variables were not mutually exclusive. Each group member created their own model using a different task. After the model was created, the group members tried using their model on the other two tasks, and the entire, pre-separated data set to identify which methods worked best for each target diagnosis.

3.1 APPROACH TASK

After the approach task data was separated, the CSV file was examined to explore the dimensionality of the features. For the target variables, all non-zero entries were changed to 1. Once it was determined that the data was high-dimensional, the data was scaled, and a correlation matrix was generated as seen in *Figure 3*.

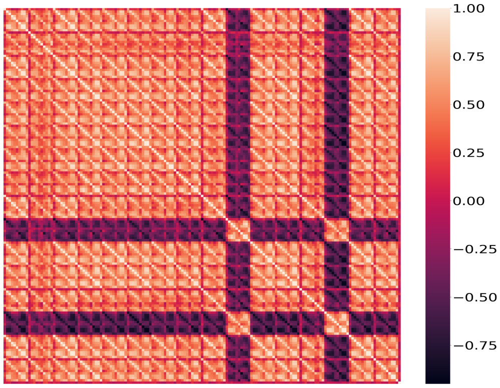


Figure 3: Correlation matrix

After examining the correlation matrix, principal component analysis (PCA) was used in order to reduce dimensions and redundant information from the highly correlated features. Next, a scree plot was used to determine the number of principal components required to contain at least 80% of the cumulative explained variance. This can be seen in *Figure 4*.

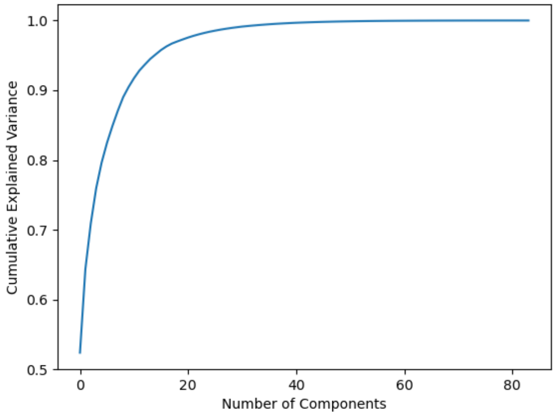


Figure 4: Scree plot

From this scree plot, it was determined that at least 8 principal components would be required. After that, a data frame was created to contain the desired number of principal components, the target variables, and a column was engineered to contain a unique number for each of the possible permutations of the target variables. This engineer information was used to generate the biplot seen in *Figure 6.* Finally, several models were used in an attempt to find one which would yield the greatest AUC values.

Five different models were used, including a logistic regression, neural network, decision tree, random forest, and extra trees. The logistic regression was chosen in case the features and target variables had somewhat of a linear relationship. It can also be regularized using L1 and L2 penalties, which can help prevent overfitting. The neural network was used because it is flexible, and can model non-linear relationships better than a logistic regression. A decision tree was used because it is good at determining important features by recursively splitting the data based on the feature that most aptly separates the target classes. A random forest was used because it uses several decision trees, which helps to improve accuracy. It can handle high-dimensional data well by evaluating split quality at each node of its decision trees, and then focusing on the features that provide the most information, although it was used on the principal components and not the original high-dimensional data set. Extra trees were used for their similarity to random forests. The main difference is that they use a different method for selecting the splits at each of its decision tree nodes, but it is still useful and reduces overfitting in a similar way to random forests.

For each model, the following process was performed: An 80-20 train-test split was used, with 80% of the actual data being used for training. Additionally, synthetic data generated by SMOTE was included in the training set. The test data was 20% of the actual data. The model was fit, and a random search was performed to determine its hyper-parameters. The model was then tested. For each model, these steps were performed thirty times using different random seeds to generate the synthetic data. The resulting performance metrics were averaged over the 30 runs. These metrics included the minimum, mean, and maximum values for accuracy (ACC), AUC, TPR, and FPR.

3.2 SPEECH TASK

The speech task data to work with consists of 264 columns of features and only 84 instances to train and test off of. After splitting into training and test sets, there were only about 60 instances to train models and about 20 instances to test models on. With 264 features, there was an assumption to not generate more features but to generate new target columns. The new target columns applied a mask to the original ADHD target column where the original column had a multiclass classification method for what type of ADHD the child was diagnosed with. The mask applied to this column binarized values to whether or not the child was diagnosed with ADHD or not. The other target column generated was a column combining the diagnoses of internalizing disorders (depression and anxiety) into a column where a value was a 1 if they had depression or anxiety and a 0 if they did not have either.

The general pipeline that was created is described next. With there being a plethora of features, there was not an emphasis on making new features but understanding the importance of the features already in the data. To start, uniform scaling of the features was performed to ensure that features that appeared to be different did not become weighted as important to the models. Next, there is a piece of code that gives the option to over sample the training data using SMOTE, which uses K-nearest neighbors to create synthetic data. Recursive feature elimination through cross-validation was implemented to cut down the amount of features that a model trained off of. Next, a grid search of hyper parameters using cross-validation was utilized to iterate over all combinations of a specified parameter grid to find the best possible parameters. After that, the option to over sample the training data was placed in the pipeline to be able to see if oversampling helped or hurt the output results. There is also an option to take the training data and split and fit the model to allow the model to train on as much data as possible.

The inputs to the pipeline that includes all the above takes in the training data, target data, the maximum number of features the model can train on, and the number of cross-validation folds which is being applied wherever cross-validation is applied in the pipeline. These were made as the inputs to the pipeline to have ways to combat over and underfitting whenever it was thought to occur. The options to execute oversampling and more splitting and fitting of training data is something that must be changed in the pipeline.

3.3 BUBBLES TASK

The Bubbles Task consisted of data collected on 16 attributes (such as sum, variance, and standard deviation, median, ect..) of a child's movement (acceleration and gyroscopic) behavior for every 15 seconds of the 3-minute task. This results in 192 features in the bubbles dataset, with 16 groups of time-series related features. Features were normalized using scikit learn min-max scaler for feature transformation. Linear, logistic, elastic net, and extra trees models were initially run on all 192 columns to determine which method showed the highest baseline accuracy and AUC scores. Two-fold stratified cross validation was used to split and train the dataset and the average score values between the two crosses were calculated for the final output. It was found that the extra trees model performed the best across all target diagnosis. As mentioned above, all ADHD diagnoses were collapsed into a 1-0 binary identifier to match anxiety and depression.

Next, two methods were used to reduce the number of features in the dataset to improve model performance. First, after noticing the 16 time-series related feature types, constructing new features of higher importance was attempted by comparing the linearity (r-squared), polynomial trends (root mean squared error) from 2 to 4, and amplitude across each of the 16 feature groups. This resulted in less than 64 features that were then used as the input for the extra trees, linear, and logistic models. Feature importance was run using the extra trees model separately on the time-series constructed feature set and for the original 192. The top 10 features were selected for each anxiety, depression, and adhd and re-run through extra trees to calculate accuracy and AUC. Comparisons were made between the constructed time series data and the original dataset. It was found that the extra trees model performed the best on the original set, so the engineered features were abandoned.

At this point the extra trees model on the original data was found to be the most accurate of the model. Trial and error methods were used to determine the number of most important features to use in the model. To improve model true positive rates, a gradient booster was implemented using the extra trees model as a base. Finally, fine tuning of parameters for both the extra trees and gradient booster were determined using randomized grid searches. In all, the final model for the bubbles task found to perform the best was a gradient boosted extra trees model with grid search, independently using the top 10 features for each diagnosis with two-fold cross validation.

4. RESULTS

Below are plots and tables that display our results for each of our respective tasks.

4.1 PREDICTIVE MODELS

The models used in each pipeline differ but between all the pipelines, the following models are utilized: Logistic Regression (LR), Random Forest Classifier (RF), Extra Trees Classifier (ET), Multilayer Perceptron (MLP), and Support Vector Machines (SVM).

4.1.1 APPROACH TASK

It was found that the model with the greatest average AUC over 30 runs was the random forest. The model was used to predict if children would have depression, anxiety, or adhd using 160 features. The performance metrics generated for each model resulted from 30 runs of the model on each of the target variables. All performance metrics for each model and disorder can be seen in Table 1.

Table 1: Performance Metrics of all models used on approach task data set

Model	Target Diagnosis	Accuracy Report: Min, Avg, Max	AUC: Min, Avg, Max	TPR: Min, Avg, Max	FPR: Min, Avg, Max
Logistic Regression	Depression	0.000, 0.261, 0.706	0.000, 0.000, 0.000	0.250, 0.250, 0.250	0.250, 0.250, 0.250
Logistic Regression	Anxiety	0.176, 0.300, 0.471	0.212, 0.401, 0.577	0.312, 0.441, 0.556	0.312, 0.441, 0.556
Logistic Regression	ADHD	0.118, 0.241, 0.765	0.767, 0.818, 0.867	0.500, 0.583, 0.600	0.500, 0.583, 0.600
Neural Network	Depression	0.647, 0.788, 0.882	0.000, 0.069, 0.438	0.250, 0.365, 0.400	0.250, 0.365, 0.400
Neural Network	Anxiety	0.353, 0.475, 0.647	0.077, 0.262, 0.519	0.250, 0.391, 0.556	0.250, 0.391, 0.556
Neural Network	ADHD	0.529, 0.733, 0.824	0.233, 0.479, 0.733	0.333, 0.428, 0.500	0.333, 0.428, 0.500
Decision Tree	Depression	0.353, 0.775, 0.941	0.125, 0.558, 0.969	0.167, 0.405, 0.857	0.167, 0.405, 0.857
Decision Tree	Anxiety	0.176, 0.498, 0.765	0.212, 0.473, 0.683	0.292, 0.475, 0.667	0.292, 0.475, 0.667
Decision Tree	ADHD	0.412, 0.706, 0.824	0.267, 0.549, 0.833	0.167, 0.407, 0.600	0.167, 0.407, 0.600
Random Forest	Depression	0.824, 0.884, 0.941	0.031, 0.636, 0.938	0.167, 0.478, 0.714	0.167, 0.478, 0.714
Random Forest	Anxiety	0.412, 0.553, 0.647	0.202, 0.391, 0.596	0.312, 0.414, 0.500	0.312, 0.414, 0.500
Random Forest	ADHD	0.647, 0.747, 0.941	0.150, 0.560, 0.883	0.278, 0.457, 0.714	0.278, 0.457, 0.714
Extra Trees	Depression	0.706, 0.875, 0.882	0.031, 0.470, 0.875	0.143, 0.443, 0.667	0.143, 0.443, 0.667
Extra Trees	Anxiety	0.471, 0.578, 0.765	0.163, 0.449, 0.673	0.321, 0.446, 0.607	0.321, 0.446, 0.607
Extra Trees	ADHD	0.647, 0.739, 0.882	0.500, 0.627, 0.733	0.333, 0.481, 0.625	0.333, 0.481, 0.625

The average AUC values of each model can be seen in Table 2. An element-wise average was calculated using the AUC values of each target diagnosis for each model. The resulting 30 AUC values were averaged to yield a single average AUC for each model.

Table 2: Average AUC of models used on approach task data set

Model	**Avg AUC**
Decision Tree	0.526766
Extra Trees	0.515166
Logistic Regression	0.406140
Neural Network	0.269833
Random Forest	0.529161

For each pairwise combination of the used models, the AUC values of the 30 runs averaged by target diagnosis were used for a Mann-Whitney U Test. This non-parametric test was used because no evidence was obtained that suggested that the data was normally distributed. The AUC values were the primary metric for evaluating model performance. It was chosen because the data set is imbalanced, which may have had a significant effect on the model's ability to distinguish between classes. It is also notable that this study pertains to medical diagnoses where incorrect classification can have significant consequences, so understanding a model’s ability to distinguish between classes is very important here. The results of the test can be seen in Table 3. Table 2 shows that the Random Forest obtained the greatest average AUC value while Table 3 shows that it did not perform significantly better than the Decision Tree or Extra Trees models. In fact, for every comparison that compared these three models to each other, the p-value was insignificant.

Table 3*:* Mann-Whitney U Test results with significance level of 0.05

Model 1	**Model 2**	**p-value**	**Significance**	**Result**
Decision Tree	Extra Trees	6.308763e-01	Not significant	
Decision Tree	Logistic Regression	1.723986e-07	Significant	Decision Tree > Logistic Regression
Decision Tree	Neural Network	3.019859e-11	Significant	Decision Tree > Neural Network
Decision Tree	Random Forest	8.533817e-01	Not significant	
Extra Trees	Logistic Regression	3.561918e-06	Significant	Extra Trees > Logistic Regression
Extra Trees	Neural Network	1.093670e-10	Significant	Extra Trees > Neural Network
Extra Trees	Random Forest	6.100076e-01	Not significant	
Logistic Regression	Neural Network	5.050032e-10	Significant	Logistic Regression > Neural Network
Logistic Regression	Random Forest	4.431012e-07	Significant	Random Forest > Logistic Regression
Neural Network	Random Forest	4.975166e-11	Significant	Random Forest > Neural Network

To delve deeper into the Random Forest model, its ROC curves can be seen in *Figure 5*.

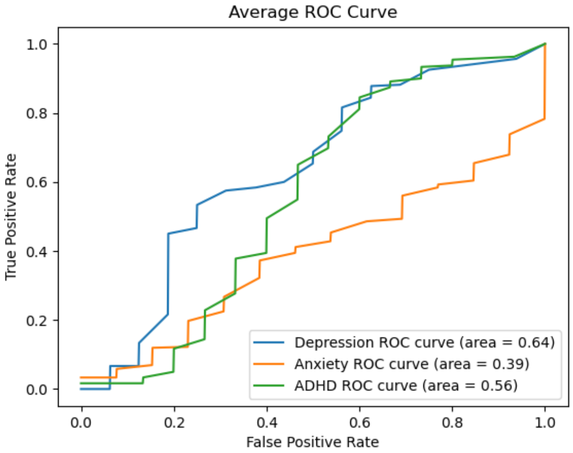


Figure 5: Average Roc Curve of Random Forest

From the ROC curve, we can see that the model performed particularly poorly at distinguishing the positive and negative diagnoses for anxiety. It appears that the model was better at separating the positive and negative diagnoses for depression and ADHD.

Additionally, a biplot was generated to determine if there was any clear separation in how each of the possible permutations of diagnoses related to the first two principal components. The plot can be seen in *Figure 5*, and shows no clear separation in how the data relate to the first two principal components.

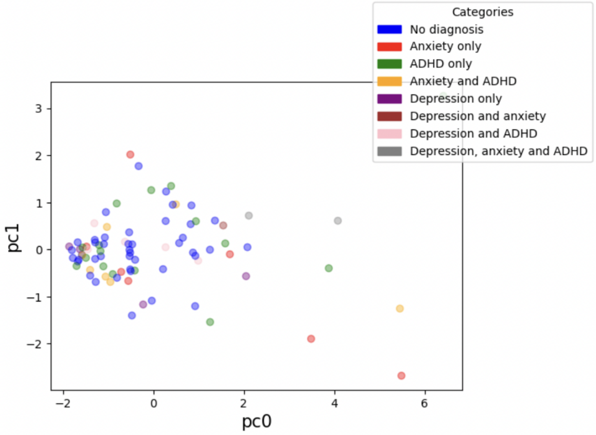


Figure 6: Biplot of first two principal components of PCA

When comparing the results of Table 1 to those obtained by the M-Sense group in *Figure 2*, it seems that the models used for the approach task in this project were not able to distinguish between positive and negative classes as well as the models used in the M-Sense group.

4.1.2 SPEECH TASK

The pipeline for this task was generalized to take data from any task or even all 616 features. Mentioned in the methods section for the speech task, there are variations for how the model can process the data and output results. To revisit, there are options for the maximum number of features the model will train on, how many folds for cross-validation, option to oversample the training data using SMOTE, and the option to take the training data and fit your model to it through cross-validation to give the model more data to learn on. For this section, the necessary information is the diagnosis, input data, the model, amount of cross-validation fold, max number of features, if oversampling was used, and the metrics outputted from all these inputs. Because of the formatting of this paper, a table with all this information would be incomprehensible, so there will be a description of the conditions of the reported outputs.

For output number 1, the amount of cross-validation fold was 10, the max number of features was 20, and oversampling was not used. For output number 2, the amount of cross-validation folds was 10, the max number of features was 30, and oversampling was performed. For output 3, the amount of cross-validation folds was 10, the max number of features was 15, and oversampling was not performed. For output 4, the amount of cross-validation folds was 10, the max number of features was 20, and oversampling was not performed.

Table 4*:* Table showing the top results of the pipeline made for the speech task.

Output num	Diagnosis	Feats	Model	ACC	ROC AUC	TPR	FPR
1	ADHD	Speech	RF	.885	.638	.667	.087
2	ADHD	Speech	MLP	.654	.665	.714	.368
3	DEP& ANX	All	ET	.692	.657	.54	.2
4	ANX	Bubbles	RF	.692	.661	.571	.263

4.1.3 BUBBLES TASK

Feature importance ranged from 0.0172 to 0.0005, showing little importance for any one feature when all 192 features are involved. The top 20 most important features for anxiety using the extra trees task are shown in figure 7. For each run of the model, between 7 and 15 features broke the threshold of 0.01 importance. Trial and error revealed anywhere from 8-12 of these top features produced the best results, therefore the top 10 features were used to standardize for the final model. A greater number of cross validation increased accuracy but decreased true positive rates, so 2-fold stratified cross validation was used. The addition of the gradient booster, using the extra trees as its base model, greatly increased the predictive capacity of the target diagnosis anxiety by approximately +5% in both accuracy and AUC scores, and +20% increase in true positive rates for anxiety (depression and ADHD showed marginal increases). The final results of the Bubble's Task are depicted in Table 5

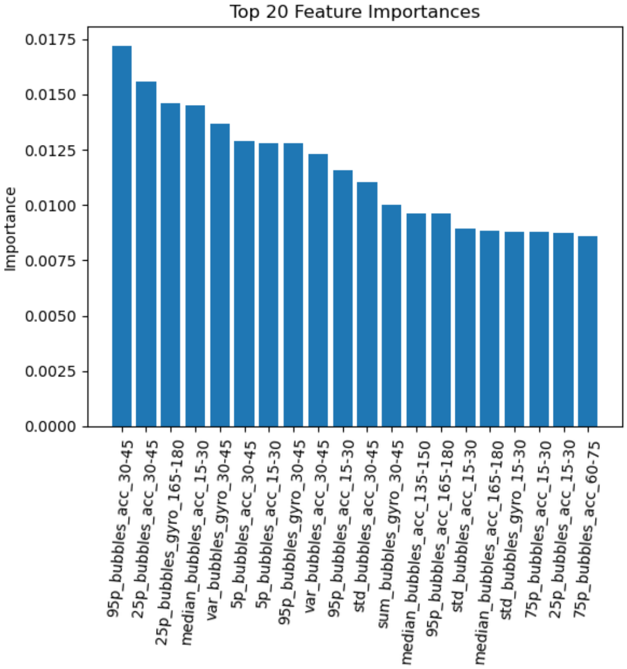


Figure 7: Top 20 features identified through extra trees model for target diagnosis:Anxiety

We found that this model performed very well in its ability to accurately predict anxiety diagnosis with an accuracy of 75%, AUC of 83%, and true positive rate of 63%, and false positive rate of 15%. These values are an improvement from the previous predictive capacity of Loftness et all 2023 for diagnosing anxiety. However, the model performed poorly in its performance for depression and ADHD, falling well below the previous benchmark. This model, extra trees with boosting, was later run on the speech and approach tasks, as well as the full dataset to test its predictive capacity on the other datasets. However, the results for anxiety were far below that of the bubbles dataset alone, and little to no improvement was found in the other target diagnosis.

Table 5: Boosted Extra Trees model on Bubbles Task

Diagnosis	Mean ACC	Mean AUC	Mean TP	Mean FP
Anxiety	0.750	0.832	0.633	0.148
Depression	0.774	0.427	0.083	0.124
ADHD	0.714	0.683	0.250	0.163

5. DISCUSSION

5.1 APPROACH TASK

In the order in which the results were presented, they will be discussed. First, the models can be ranked by average AUC in the following order: Random Forest, Decision Tree, Extra Trees, Logistic Regression, Neural Network. While the Random Forest model performed better than the others, its p-value was not significant for every pairwise comparison under a Mann-Whitney U test. Notably, any pairwise comparison between the Random Forest, Decision Tree, and Extra Trees resulted in an insignificant p-value. Some possible factors that influenced this result might include a low sample size, the hyperparameters of the models, and the similarities of the models. All of the models are based on decision trees, and the random forest and extra trees models use a bagging technique called "bootstrap aggregating" in order to generate multiple training sets by randomly sampling the data to train decision trees. The similarity in how these models work may have contributed to how similar their AUC values were. The AUC data suggested that there was a statistically significant difference between the logistic regression and neural network with the random forest, decision tree, and extra trees models, with the logistic regression and neural network contextually having the lower mean AUC value. The logistic regression may have performed worse than the tree-based models because the data may not have been linearly related to the target variables. The tree-based models are better at interpreting non-linearly related data. The p-value for the comparison between the logistic regression and the neural network was significant, with the neural network having the contextually lower mean AUC value. While the linear regression may have been a poor model choice for the data set, it could have performed better than the neural network, its superior performance might be attributed to a lack of computational resources allotted to the neural network.

Limitations: The lack of data and imbalance of classes was a major limitation in this project. It is indicated from the AUC values in *Table 1* that the tree-based models struggled to distinguish between positive and negative classes of anxiety, which could have been impacted by the small size of the data set. Additionally, the imbalanced distribution of diagnoses in the data set led to the use of SMOTE to generate synthetic data via the training set, which may not have been a true representation of the actual data. Additionally, PCA was used to reduce the dimensionality of the data set, which may have obscured information that would have helped to distinguish cases of anxiety from those without it. Also, the computational resources allotted to determine the hyperparameters of the models may have not been sufficient to reach the models' potential performances.

5.2 SPEECH TASK

The results for the speech task indicate that there are no good results for the depression diagnosis. Even with the ability to vary the combination of settings to run the model, dramatic overfitting or underfitting seems to be taking place. The output for all models is either exactly perfect at classifying or classifies everything as not having depression. The reason for this behavior with the depression target is unclear and deserves further inquiry.

Besides that, the other results are notable in that they are better than the benchmark in some way. Output 1 is an improvement in accuracy and false positive rate as well as true positive rate for the ADHD diagnosis. Output 2 is roughly similar to the results for ADHD in the benchmark table but tends toward a higher true positive rate which is why it is included. Output 3 is included to show that using all the features without pruning can potentially predict internalizing disorders (depression and anxiety). The true positive rate is not very high but still higher than one of the true positive rates in the benchmark table. Output 4 has better results on almost all fronts and classifies anxiety rather well.

Having a limited amount of data and imbalance classes makes this problem interesting in the following way. If the training set happens to contain a majority of the positive diagnosis class, it should be more proficient in classifying a positive diagnosis. However, since the majority of the positive diagnosis class is in training and not test set, the model will not have much of an opportunity to classify positive diagnoses when it counts towards the actual model accuracy. The same issue arises in the reverse of this problem, where the majority of the positive diagnosis is in the test set and not in the training set. In this case, the model will not have a good idea as to different ways a positive diagnosis might appear, and then mischaracterize them in the test set, bringing down all the metrics.

5.3 BUBBLES TASK

The results of the boosted extra trees model on the bubbles task revealed surprising predictive capacity for anxiety, yet not the other tasks. This proves that machine learning models have the potential to be a useful tool, in addition to traditional methods, of diagnosing individuals with internalized ADHD disorders. It remains unclear to us why the bubbles task performed to such a higher degree over the other tasks, particularly on identifying anxiety disorders. Additionally, it was a surprise the time series engineered features did not show improvement on the data, and it may be worth looking back into extracting other measurements of time series trends beyond what was analyzed r-squared, rmse, and amplitudes.

Limitations: The most clear limitation of this study is the lack of individuals to train these models on. This is most likely due to the small sample size of individuals in the study who were diagnosed with depression or ADHD. Anxiety had the largest number of individuals in the dataset and this may be why the model's performance was better, indicative of the potential for this kind of research given a larger number of participants for the other diagnosis. Additionally, when this model was later run on the other tasks and the entire dataset, providing worse results than bubbles alone. This brings up a few concerns. The fact that the model, even for anxiety, performed worse when given the full dataset brings into question the methods used in feature selection. The more features included, the less important each one becomes in the extra trees model, and perhaps overloading the model with over 600 features proved too complicated to identify which features actually had the most impact when this model was run on the entire dataset rather than bubbles alone. This model was later run on the speech, approach, and full dataset. predictive capacity for anxiety was lower in all other cases than the bubbles task alone, and depression and ADHD did not show improved performance from the other tasks.

6. FUTURE WORK

An obvious way to improve the results of the study would be to have more data to work with and train models with. This cannot be helped but it would most likely improve the results. This lack of data is what makes this project difficult and fun to work with. There was much attention to how data leakage could occur and biases in how to train the data. Related to this, one other piece of future work would be to conduct this project but to split the data by gender or group together children of the same age. This might not be the most helpful because this would dramatically reduce the pool of data to then split into a training and test set for the models.

One other way to continue this would be to conduct statistical tests with the models to determine what models worked the best for what input features and to also find what hyperparameters worked for said models. This project was a good survey showing that there is potential to improve the benchmark results but with the amount of time, it was only a survey. It would be satisfying to uncover consistent methods that work in analyzing this data.

Specifically for the speech task pipeline, if there was more time and better computational resources, iterating through values of cross-validation folds and maximum number of features for each feature set could prove valuable in finding the best hyperparameters to inspect every feature set. A given run, going through depression, anxiety, ADHD, and then depression and anxiety targets, takes anywhere from 30 minutes to an hour and a half. Iterating through the cross-validation and maximum number of features for each feature set could have easily taken at least 20 hours and then doing that four times for each target for a total of at least 80 hours.

7. CONCLUSION

Based on the AUC values obtained for the approach task, it appears that the tree-based models that were used for the approach task were either worse or similar in performance to those used by the M-Sense group. While several factors may have contributed to the models' lower performance such as the imbalance distribution of diagnoses and the use of PCA, it is clear that further research is needed to improve the models' ability to distinguish between positive and negative diagnoses for approach data.

The pipeline created for the speech task, with limited time and resources, was able to produce results shown in Table 4 that rivaled or improved upon the results of the benchmark. The results found could only be scratching the surface for what the pipeline is capable of finding. Models found by this pipeline were able to find results that improved the prediction of almost all the diagnoses but depression. With the amount of tinkering options in the pipeline, there are certainly more models to be found that either rival or improve from the benchmark.

The methods used for the bubbles task showed promise for a tool to help diagnose anxiety, beating the benchmark set by the M-Sense group for this diagnosis. This model, however, performed far below the benchmark for depression and ADHD, suggesting that different diagnoses may require alternative methods to best optimize these modeling techniques for the diagnosis of different disorders. Regardless, these models show promise for future screening of disorders in children, given more data can be collected to better train the models on.

8. REFERENCES

[1] Faedda GL, et al. Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. *J Child Psychol Psychiatry*. 2016;57(6):706–16. doi: 10.1111/jcpp.12520.

[2] H. L. Egger and A. Angold, "Common emotional and behavioral disorders in preschool children: presentation, nosology, and epidemiology," *J Child Psychol Psychiatry*, vol. 47, no.3–4, pp. 313–337, Apr. 2006, doi: 10.1111/j.1469-7610.2006.01618.x.

[3] Loftness, B. C., et al. (2023). The ChAMP App: A Scalable mHealth Technology for Detecting Digital Phenotypes of Early Childhood Mental Health. medRxiv. https://doi.org/10.1101/2023.01.19.23284753

[4] Loftness, B. C., et al. (2023). Toward Digital Phenotypes of Early Childhood Mental Health via Unsupervised and Supervised Machine Learning. medRxiv. https://doi.org/10.1101/2023.02.24.23286417

[5] M. Tandon, E. Cardeli, and J. Luby, "Internalizing Disorders in Early Childhood: A Review of Depressive and Anxiety Disorders," *Child and Adolescent Psychiatric Clinics of North America*, vol. 18, no.3, pp. 593–610, Jul. 2009, doi: 10.1016/j.chc.2009.03.004.

[6] N. R. Towe-Goodman, et al., "Perceived Family Impact of Preschool Anxiety Disorders," *J Am Acad Child Adolesc Psychiatry*, vol. 53, no.4, pp. 437–446, Apr. 2014, doi: 10.1016/j.jaac.2013.12.017.