Predicting Efficacy of Therapeutic Services for Autism Spectrum Disorder Using Scientific Workflows

Fahima Bhuyan, Shiyong Lu, Ishtiaq Ahmed
Department of Computer Science
Wayne State University
Detroit, MI
{*fahima.amin; shiyong; ishtiaq*}*@wayne.edu*

Jia Zhang
Department of Electrical and Computer Engineering
Carnegie Mellon University
Mountain View, CA
jia.zhang@sv.cmu.edu

***Abstract*—Early intervention in autism, although deemed as essential, has high variance in the outcome attained, partially due to complex interaction between multitude of factors and variables involved, and the lack of systematic study to untangle their influences in the outcome. Therefore, pairing set of interventions with an individual children to cater for their need remains highly challenging. From the perspective of parents, unknown factors emanate from their unfamiliarity with what interventions are out there and why. From the perspective of caregivers, it is critical to understand unique attributes of the individual children develop over time. There is a scarcity of exploration of interactions between attributes specific to a child, family characteristics and therapeutic, medical and educational services. In this research, we aim to bridge the gap. In this study, we identify predictive features pertaining to each individual child and how they interact responding to different interventions and services. We have studied temporal data and model improvement/regression outcomes at different timestamped milestones and overlayed a model to aid parents and caregivers in coming up with pragmatic intervention plan. We propose a scientific workflow to automate the modeling process and rely on DATAVIEW to guarantee computational reproducibility and data fidelity. We use data collected by SFARI dataset for evaluation. To the best of our knowledge, this is first-time amalgamation between the Autism Health informatics community and the Workflow community; and this is the first-time study that combines prediction methods applied on Autism Spectrum Disorder (ASD) Phenotype data to provide guidance to parents and caregivers.**

***Keywords*-autism; data-mining; workflow management system; DATAVIEW;**

I. Introduction

In recent years, we are noticing a sudden spike in Autism Spectrum Disorder (ASD) in the US population. With Autism in the rise, there lacks retrospective study in management and alignment of behavior intervention and how it affects manifestation of ASD symptoms. As a result, parents and community at large are finding it challenging to personalize treatment plans for kids with ASD. Meanwhile, there exists a flurry of anecdotal evidences and data recorded by parents, caregivers and therapists. These data, if mined retrospectively, may lead to better understanding of how appropriate goals, given a child's traits and needs, could potentially mitigate autistic behavior and result in overall improvement. In this research project, we aim to find correlation and causation between improvements in individuals in the spectrum, and how appropriate behavior and educational goals can lead to betterment of symptoms. As this study involves numerous variables, we will focus on mining information pertaining to phenotype data including updated medical, developmental and educational history.

Heterogeneity in data collection and interpretation in treating autism spectrum disorder is one of the most fundamental challenges in treating autism and creating pragmatic intervention plan. Since manifestation of autism is caused by hundreds of genes [1], several neuro-cognitive mechanism, variance in phenotypical features also lie in a wide spectrum. Different areas in this phenotype spectrum warrant uniquely catered intervention plan.

Analyzing collective data suggests that early intervention in treating autism spectrum disorder is highly effective in improving social, adaptive and communication skills of affected kids. However, drilling down into the data suggests that individual response to early intervention is highly variable. Some children respond with substantial improvement, while others with marginal or no improvement at all. Hence, a blanket statement on the efficacy of early intervention on an aggregated level, while true, does not warrant improvement on a more individual basis.

Autism spectrum disorder, from the standpoint of parents and caregivers, can sometimes be construed as an enigma. There are potentially numerous reasons why ASD is deemed rather perplexing. Among others, lack of data, no clear treatment plan other than Adaptive Behavior Analysis, and lack of fine-grained studies are a few salient reasons. A common question that parents keep asking experts is how they can help their children make progress, e.g., speech therapy, ABA and occupational therapy hours, or creating an appropriate goal plan. However, the treatment plan created for a particular child not only depends on the behaviors that the child exhibits, but also varies because of the quality of the therapy provided and the resources available. Thus, it has become imperative to aid families in coming up

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

with appropriate set of goals, generated from a data-driven standpoint and eliminating individual biases. For example, knowing that providing more mainstreamed education might help a child make progress, irrespective of whether mainstreaming opportunities are available at the institution, might help parents make decision about their child's placement.

In this study we aim to focus on the following four aspects:

- We propose to systematically create a workflow, based on individual and anonymized data, to unveil the most effective mode of symptom status for a child.
- We propose a scientific workflow to automate the modeling process and rely on DATAVIEW to guarantee computational reproducibility and data fidelity.
- We propose to mine historical data to identify correlation between temporal improvement in ASD symptoms.
- We propose to develop a platform to share anonymized data of temporal aspects of ASD traits, which can improve statistical studies in identifying how the traits can be managed in a data-driven and a practical manner.

II. Background

Autism, a recent epidemic of a medical condition requires attention from medical, data science, and behavioral communities for analyzing cause and trend, as well as for predicting future outcomes. Autism is a neuro-developmental disorder that impairs natural development, causes challenges in emotional interaction, social communication, sensory processing etc. Autism has a wide range of symptoms that is why it is referred to as Autism Spectrum Disorder (ASD). In early 1940s, the condition was named as "Autism" and "Asperger Syndrome" by Leo Kanner and Hans Asperger, respectively [2]. Based on the data on National Institute of Mental Health (NIMH), in the 1970s the ASD rate was 1 in 10,000. In the late 1995, the rate increased to 1 in 1,000. In 1999 the rate became 1 in 500. In 2001, the rate was 1 in 250. In 2005, the rate was 1 in 166. In 2007 the rate was revised to 1 in 150. In 2009, the rate rose to 1 in 90 [2]. Nowadays, ASD affects approximately 0.5-0.6% of the population [3].

The rapid growth in ASD rate warrants a reason for people to study environmental factors and adopt data-driven approaches to analyze its causes and symptoms. Due to proliferation of research in this direction, there appeared a large pool of data to where people can analyze using data mining techniques. Moreover, advanced techniques of storage and mining of big data are leveraged to help understand ASD better.

In the process of promoting research and collecting data, Dr. Bernard Rimland [2] founded the Autism Research Institute (ARI) in San Diego, CA in 1967. Their core mission is to identify the cause of autism and to evaluate treatment efficacy. To date, they have collected survey data from over 40,000 parents of children with ASD throughout the world [2].

To better understand this epidemic medical condition, there are different organizations collecting confidential data from families with ASD children. Each year, more families become willing to store their data either for the sake of predicting the chances of their siblings having same conditions, or with hope of knowing the cause of their childen. National Database of Autism Research[1] has archived the Phenotype data collected from families and professionals. Based on the available concepts from NDAR, we have generated graphical representation for depicting the wide range of features associated with Autism Spectrum Disorder (ASD). Fig. 1 shows a highly brief hierarchy of the concepts.

III. Related Work

Recently, few researches conducted based on amalgamation of machine learning, data mining and autism data analysis. This is a comparatively new area, where a behavioral and neurological problem is analyzed in perspective of machine learning and data mining techniques. The research in this area mostly gained attention since 2009. Unfortunately, there lacks a uniform platform for addressing all the issues. Here we will discuss some of the related work done in the area of ASD based on machine learning and data mining techniques.

An automatic alert system, Autistic Child Sensor and Assistant System (ACSA), has been developed for autistic children and their families for protecting the child from overstimulating environments, incidents and injuries, using wireless sensor network [4]. The system aims for detecting and processing autistic movements based on machine learning algorithms. ACSA works on three different components: an ACSA wearable sensory device is worn by autistic children on appropriate locations of their body determined by physician or family; an ACSA parent application is installed on parents' smart devices; machine learning algorithms are used for actively recognizing and accurately detecting children's gestures and motions.

For capturing discriminative eye movement patterns, a prediction system has been developed for discovering the latent patterns based on eye movements from the sequentially recorded images [5]. There are other studies that show how machine learning process is used to optimize the diagnosis process, by tracking eye movements of ASD children [6], [7], [8]. Compared to typically developing individuals, ASD children and adults show reduced visual attention to faces [9]. Work in [10], [11] shows the evidence of different eye movements by ASD individuals when scanning faces.

There is Autism and Emotion sub-challenges introduced by INTERSPEECH 2013 computational Paralinguistics Challenge [12]. They provide the results based on integration of multiple well-known machine learning techniques, including support vector machine, deep neural network,

[1] https://ndar.nih.gov/

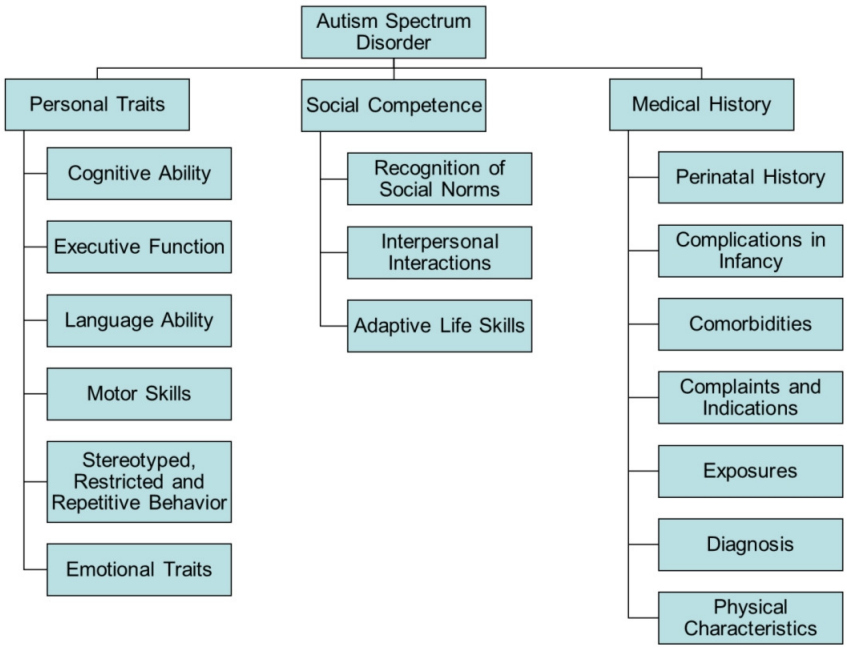

Figure 1: Overall view of Autism Spectrum Disorder.

weighted discrete K-nearest neighbor; and the ASM (Acoustic segment model) approach.

A human-robot interaction technique has been used in the direction of future approaches for modeling the behaviors of children with autism [13]. There are case studies and anecdotal evidences that show children with autism exhibit improved social behavior with robots. Two machine learning techniques are applied: Conditional Random Fields(CRF) to both classify and segment in time series data generated by experiments, and decision tree $C4.5$ algorithm in Weka tool to predict vocalization. Some research groups also have used robots for children with autism [14], [15], [16].

The approach in [2] provides a better understanding of ASD, by selecting features for identifying subtypes of autism to determine the effectiveness of clustering for further classification purposes.

Another research has been done on mining ASD data based on twitter information [3]. It is the first paper focusing on using social networking for data-mining information related to ASD. Their investigation gives more concerns, practices and generally more topics in conversation with people interested in ASD and motivating further work towards the direction.

Other data mining technologies have also been used for investigating feature selection in gene expression [1]. They used classification method for selecting genes and gene sequences between ASD and healthy cases.

AMP (Autism Management Platform) [17] is a mobile application and intelligent web interface for capturing, analyzing and managing data associated with diagnosis and treatment of ASD. Though continuous data management remains a challenge, this analytic platform aggregates and mines patient data in real-time and gives relevant feedback based on automatically learned data by filtering preferences over time.

IV. Problem statement

In this research problem, we aim to develop a scientific workflow to analyze phenotype data and find attributes and relations in these features. We also intend to investigate how to identify factors for success, and antecedents for behavior and regression. Based on historical data, we aim to run time series analysis to identify factors that parents can invest in to work toward positive behavioral and educational outcome.

To date, we have not seen prediction method or mining technology that predict the trend of successful treatment based on the data in different subjects and environments (i.e., general Ed. or special Ed.). Mining historical data may pave a way to understanding effectiveness of mainstreaming for children on the spectrum and guiding parents and school districts in making decisions.

For ASD children, one of the effective intervention methods is ABA (Applied Behavioral Analysis). Evidence shows that ABA works better than all other behavior therapies. ABA data are collected in the form of ABC - Antecedent, Behavior and Consequences. As most of ASD kids have limited language skills, ABC data gives a good understanding of their behavior, both problematic and good.

- A. Antecedent Data: It gives a good insight of antecedent of any problem behavior that triggered the

behavior.

- B. Behavior Data: Behavior that is presented by the child.

- C. Consequences: A protocol of good consequence is by which the behavior can be shaped.

There are several ways to analyze a child's data. One way could be, in our workflow management system, to analyze the childś behavioral data to analyze what kind of educational setting he is more compliant to, which essentially predicts the effective environment for the child for learning purpose.

V. Proposed Work

A. Predictors of Improvement in Treatment Response

Traits in individual children and fine-grained intervention plans can be viewed as one of many temporal instances of a bipartite graph, where some of the interactions between these two groups result in positive outcome while others result in regression or no progress. The temporal aspect of the assignment means these interactions and their characteristics alter in terms of efficacy and effectiveness with time. Having a data-driven model that answers these questions while planning early and subsequent intervention would improve the expected outcome. Many times interventions chosen for a specific child is decided by the availability of services, anecdotal evidence instead of data-driven decisions, resulting in sub-optimal outcome from the intervention. Having children enroll in less than ideal interventions causes financial burden on the families with little or no observable gain. In this work, we aim to establish a theoretical framework to mine relevant information from data to guide intervention plan catered with individuality in mind. Historically, data collected during interventions can be broken down into several groups:

- Aberrant Behavior Checklist (ABC)
- Adult Behavior Checklist for Ages 19 to 59 (ABCL)
- Autism Diagnostic Interview-Revised (ADI-R)
- Autism Diagnostic Observation Schedule (ADOS)
- Broad Autism Phenotype Questionnaire (BAPQ)
- Child Behavior Checklist for ages 6 to 18 years (CBCL)
- Repetitive Behavior Scale-Revised (RBS-R)
- Social Communication Questionnaire (SCQ-L) - Parent report
- Social Communication Questionnaire (SCQ-C) - Teacher report
- Social Responsiveness Scale (SRS)
- Vineland Adaptive Behavior Scale-II (VABS-II)

B. Methods

As seen in Fig. 2, the entire prediction problem can be modeled as a sequence learning problem, where at each timestamp the interaction between three aspects of the data set, i.e., child's attribute, family history and medical history contribute to the expected improvement. Moreover, expected outcome in the previous timestamp also influences the current outcome. Hence, the problem can be modeled with markovian assumption. With this assumption, we can essentially model it as a Conditional Random Field (CRF) problem. Once the model is trained, the viterbi lattice can be modified to explore the current set of actions that lead to expected improved outcome in the future. We explore linear chain conditional random field to model the temporal improvement in phenotype. Conditional random field and hidden markov models have been used historically in NLP problems such as part-of-speech tagging, named-entity recognition and other problems such as speech processing and they are well suited to model temporal dependency of data. Since, changes in phenotype is highly dependent on the previous timestamp as well as intervention plan that is currently in place, we further frame it as a temporal modeling problem.

$$p(l|s) = \frac{exp[\sum_{j=1}^{m}\sum_{i=1}^{n} f_j(s,i,l_i,l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m}\sum_{i=1}^{n} f_j(s,i,l'_i,l'_{i-1})]} \quad (1)$$

In Eq. 1, l stands for the sequence of labels that we are interested in and s stands for the sequence of inputs. Essentially in linear chain CRF, the problem depends on time-window sensitive features and labels from adjacent time windows. In this case, the conditional is modeled based on data. The problem can also be formulated as a hidden markov model (HMM), where the joint probability is modeled and using the bayesian rule, the conditional is inferred. As shown in Eq. 2 for HMM, the joint probability is modeled as a product of the probability of sequence of labels and the probability of emission probability. Here, we assume first order markov dependency, where the observations in current timestamp depends solely on the timestamp before and is agnostic of the features in the timestamps before that.

$$p(l,s) = p(l_1)\prod_i p(l_i|l_{i-1})p(w_i|l_i) \quad (2)$$

To understand the influence of time agnostic variables, we can also model the problem without considering temporal variables, and regress to identify expected improvement in behavior based on features observed in a specific window of time independent of other timestamps. A by-product of the analysis would be identification of attributes (i.e., intervention plans and interactions between the trifecta as stated above) that lead to a better response to a specific

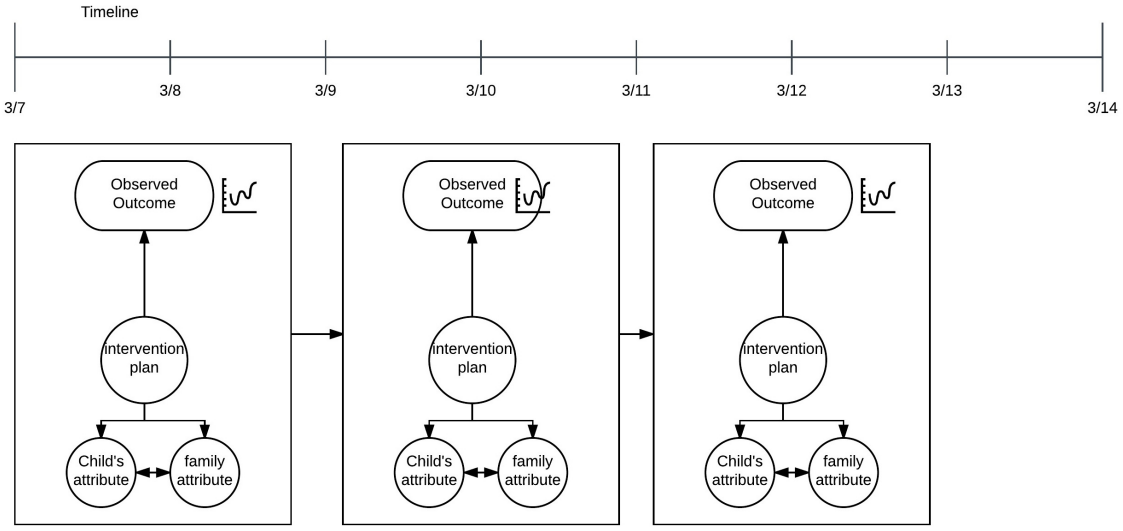

Figure 2: Treatment Improvement Predictor.

intervention plan. We can model time agnostic analysis as a regression problem, i.e., logistic or linear regression as shown in Eq. 3. Proper regularization regression analysis can also provide significance of individual features and hence can guide what interventions to pay attention to. However, since the notion of dependence in time window is lost in this case, and we model the problem by collapsing multiple windows into a single input frame, modeling variable length time window becomes challenging.

$$log(\frac{p_i}{1-p_i}) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} \quad (3)$$

However, the model, although simple, as it is modeled as a linear combination of the features, can provide important information about the features themselves.

C. Modeling Scientific Workflow

We propose to structure the entire learning problem as a scientific workflow and propose strategy to guarantee reproducibility and data fidelity. Hence, the workflow can be deemed as portable and can be retrained and reused in isolation, for example, for different age groups. After data is collected, the entire process can be modeled as a scientific workflow. We can model data pertaining to children, family and intervention plan as temporal data sources in the workflow. We aim to make use of data-mining tools in DATAVIEW and augment the system with capability to handle temporal modeling techniques. Validation of the method will be done based on standard techniques, i.e., label wise precision-recall as well as anecdotal validation from parent interviews.

Let $C_t = [c_1, c_2, ...c_n]_t$ be the features related to ASD symptom of a child;
$F_t = [f_1, f_2, ...f_m]_t$ be family history and features;
$I_t = [i_1, i_2, ...i_p]_t$ be intervention plan.
We propose a model $P(C_t|c_{1....(t-1)} f_{1....(t-1)} i_{1....(t-1)})$ for linear chain CRF:

$$\frac{exp \sum_t \sum_k \lambda_k f_k(c, f, i, t, t-1)}{\sum_{l'} exp \sum_t \sum_k \lambda_k f_k(c, f, i, t, t-1)} \quad (4)$$

Table I shows the features involved for predicting treatment outcomes.

The features of the child can be further modeled in two separate categories. $c1$: i.e., portion of the feature that can be influenced by the other features; $c2$: features that are uninfluenced by the rest of the inputs as shown in Table I. If we model the problem as a linear chain CRF, we aim to predict sequence of $c1$ features over a time window. Once the model is trained, the computed lattice can be expanded to retrospectively explored to identify how the outcomes, i.e., sequence of $c1$ features, can be perturbed in the right direction.

The Table II depicts how linear chain CRF views the problem. At each timestamp, we have features from the child, family features and intervention features. Based on the data and previous windows, $c1$ is predicted for the next window.

Modeling it as a regression problem is somewhat simpler. We decided on a predefined window size, where in the example it is 4. The features of the window are fed into

Table I: Factors involved for predicting treatment outcomes.

ASD child predictable features c1	*ASD child given features c2*	*Family Features f*	*Intervention Plan i*
Denial Tolerance/time unit	Age	Father's education level	Hours of ABA/week
Elopement/time unit	Weight	Mother's education level	Speech/week
Listener Response/ time unit	IQ	Father's IQ level	Behavior Intervention Plan
Mand/time unit	Food Habit	Mother's IQ level	Nutrition supplements/ week
MLU/time unit	Geographic location	Family History of ASD from Father's side	
Throwing/time unit		Family History of ASD from Mother's side	
Hitting/time unit		Family stress level	
Dropping/time unit		Family's positive involvement	
Self InjuriousBehavior/time unit		Social support	
Property Destruction/time unit		Family's expectation about treatment	

Table II: Feature Prediction Based on Timestamp.

Timestamp 1	*Timestamp 2*	*Timestamp 3*	*Timestamp 4*
c1 c2	c1 c2	c1 c2	? c2
F	F	F	F
I	I	I	I

the algorithm, and the outcome is modeled as a linear combination of the features. This modeling technique does suffer from non-linearity in the feature space, however, that can be overcome by employing proper transformation in the feature space.

VI. Implementation and Experiments

A. DATAVIEW: A Big Data Workflow Management System

DATAVIEW is a big data workflow management system [18], that shows the feasibility of learning computational thinking in perspective of scientific workflow. We have used this workflow management system for implementing data mining techniques for predicting the outcome based on the features available. The main reason of using DATAVIEW is to give flexibility to researcher of Autism Community and also parents and caregivers not to deal with any underlying complexity of computation, and can predict or correlate between features based on given train dataset. This also gives us the platform for working on big data. The training dataset is a continuous process of growing towards big data as more datasets are available to append in training dataset and make the prediction of test data more predictable and dependable.

The DATAVIEW has many primitive workflows, called built-in workflows, developed by DATAVIEW team. After primitive constructs are developed by developers, it is comparatively easy and convenient to build executable workflows. For each built-in or primitive workflow used, one only need the required numbers of input and output.

Moreover, for giving users more flexibility, the current DATAVIEW has integrated Dropbox feature, so that users can provide any big data file and computation tools to analyze those data and drag-drop the built-in constructs provided by DATAVIEW team and save the result in output which also can be accessible in a Dropbox folder. In this way, end users do not have to deal with the underlying complexity and can easily obtain results.

Fig. 6 shows an example of our executable workflow for predicting classes based on Random Forest data mining technique.

B. SFARI Dataset

We have collaboration with the SFARI (Simons Foundation Autism Research Initiative) [19] research and gathered a wide range of data which are collected over a period of time. Here SFARI gathered raw data of 440 families where the number of individuals with diagnosis is 1,050. Here we used only the Phenotype and their followup.

In this followup study, participants completed a variety of measures which include updated medical, educational histories and developmental updates based on standardized questionnaires.

Based on raw data collected, after doing preprocessing of data and analysis, we have created three tables based on Phenotype and their followup data:

- FollowUpMedTbl,
- FollowUpEduTbl and
- FollowUpFamHisTbl.

In Fig. 4, we can see the attribute set selected for each table from three different contexts. These are all Proband data which means ASD diagnosed person.

In FollowUpMedTble, some of the salient attributes are *sscmedcodevalue*, *symptomstatus*, *type* where based on medication given, we can see the symptom status of the proband,

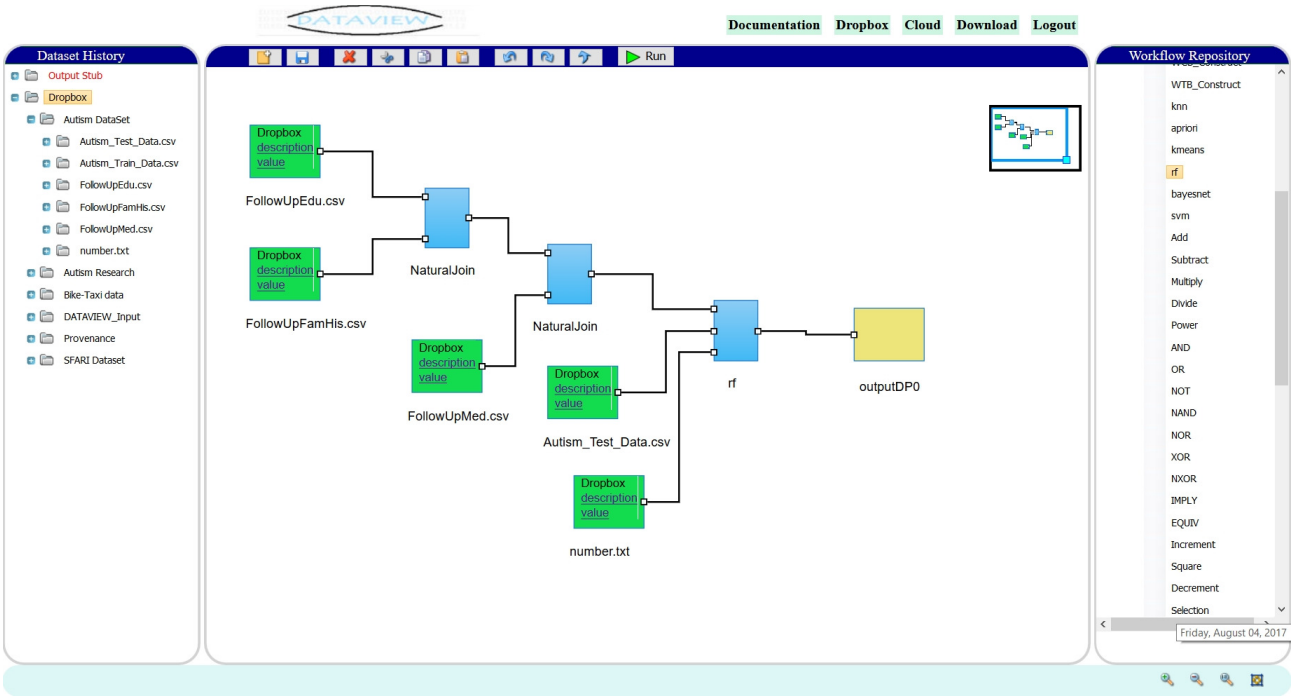

Figure 3: Running Workflow Predicting classes based on Random Forest Data Mining Technique.

i.e., symptoms worsened, mo change in symptoms, symptoms have improved for past, current or current other.

In the FollowUpEduTbl, we can see which school type they are in, i.e., Special Ed or General ed or combination of both, what is their grade level, what kind of services they are getting, how is the classroom setting, do they have any personal aide, also information about the siblings, are siblings also receiving intervention services or special ed services.

In FollowUpFamHisTbl, we have all the information about their ages, and what kind of challenges they are facing. ASD is a combination of challenges. Looking into those attributes we can see the challenges could be any one of the following: *language*, *phonological*, *developmental*, *learning*, *intellectual*, *epilepsy*, *adhd*, *ocd*, *anxiety*, *depression*, *bipolar*, *schizophrenia*, or other.

After getting all the information from three different contexts and joining all the information, we have a complete set of findings for one individual. We analyze this complete set of data using two data mining algorithms and run that in DATAVIEW.

C. Running Random Forest Algorithm in DATAVIEW

We used Random Forest as our prediction method. Random Forest is an ensemble learning method for classification, regression and other tasks by constructing multitude of decision trees at training time. This random decision forest method outputs the class that is the mode of the classes in classification or mean prediction in regression of the individual trees.

In order to depict the learning process, we delineate it with an example of decision tree. Based on these features, we aim to predict the outcome of symptom status. We train a decision tree for exposition purposes.

To run the Random Forest algorithm in DATAVIEW, we used DATAVIEWś primitive workflow, “NaturalJoin”. We joined Education History and Family History first, then used another “NaturalJoin” primitive workflow to join individual’s medical history too.

Now we use our primitive Random Forest workflow named “rf” which run the data mining algorithm Random Forest. It has three input port and one output port. Input ports are training set, test set and the number of trees we would like to generate. The final join output works as an input for the training set. For test set, we can label any attribute we want to predict. The output port will return the correct label of the test set.

For this experiment we predict, the symptom status of proband based on the training data. We show our prediction accuracy based on PR curve in Fig. 5. Here the precision recall curve is plotted for the class “Symptoms improved.” We observe similar performance for the other classes as well.

D. Running Support Vector Machine Algorithm in DATAVIEW

We ran the same experiment using Support Vector Machine. We used the same three tables and our primitive work-

```
CREATE TABLE FollowUpMedTbl(
 sfari_id VARCHAR(8),
 multiplex VARCHAR(9),
 lost_diagnosis VARCHAR(14),
 sex    VARCHAR(6) NOT NULL,
 role   VARCHAR(7) NOT NULL,
 Fcode   VARCHAR(10) NOT NULL,
 sscmedFcodedvalue VARCHAR(32) NOT NULL,
 Ftype   VARCHAR(13) NOT NULL,
 symptomstatus VARCHAR(22),
 age_at_eval VARCHAR(11) NOT NULL,
 measure_Fcode VARCHAR(27) NOT NULL,
 measure_Ftype_revision VARCHAR(27) NOT
NULL,
 study   VARCHAR(13) NOT NULL
);
```

(a)

```
CREATE TABLE FollowUpEduTbl (
 sfari_id VARCHAR(8) NOT NULL PRIMARY KEY,
 multiplex VARCHAR(9),
 lost_diagnosis VARCHAR(14),
 sex    VARCHAR(6) NOT NULL,
 role   VARCHAR(7) NOT NULL,
 Fcode   VARCHAR(4) NOT NULL,
 school_type VARCHAR(60),
 grade_level VARCHAR(42),
 special_ed_services VARCHAR(19),
 special_ed_Fcode VARCHAR(35),
 classroom_setting VARCHAR(73),
 personal_aide VARCHAR(47),
 siblings_intervention_services VARCHAR(30),
 siblings_special_ed_services VARCHAR(28),
 age_at_eval VARCHAR(11) NOT NULL,
 measure_Fcode VARCHAR(12) NOT NULL,
 measure_type_revision VARCHAR(27) NOT NULL,
 study   VARCHAR(13) NOT NULL
);
```

(b)

```
CREATE TABLE FollowUpFamHisTbl(
 sfari_id VARCHAR(8) NOT NULL PRIMARY KEY,
 multiplex          VARCHAR(4),
 lost_diagnosis     VARCHAR(30),
 sex                VARCHAR(6) NOT NULL,
 role               VARCHAR(7) NOT NULL,
 Fcode              INTEGER  NOT NULL,
 relationship       VARCHAR(38) NOT NULL,
 yearofbirth        VARCHAR(4) NOT NULL,
 gender             VARCHAR(1) NOT NULL,
 asd               BIT  NOT NULL,
 scd               BIT  NOT NULL,
 language           BIT  NOT NULL,
 phonological       BIT  NOT NULL,
 developmental      BIT  NOT NULL,
 learning          BIT  NOT NULL,
 intellectual       BIT  NOT NULL,
 epilepsy           BIT  NOT NULL,
 adhd              BIT  NOT NULL,
 ocd               BIT  NOT NULL,
 anxiety            BIT  NOT NULL,
 depression         BIT  NOT NULL,
 bipolar           BIT  NOT NULL,
 schizophrenia      BIT  NOT NULL,
 other             BIT  NOT NULL,
 age_at_eval        VARCHAR(11)  NOT NULL,
 measure_Fcode      VARCHAR(30)  NOT NULL,
 measure_type_revision VARCHAR(30) NOT NULL,
 study              VARCHAR(13) NOT NULL
);
```

(c)

```
SELECT m.sfari_id, m.multiplex,
m.lost_diagnosis, m.sex, m.role, m.Fcode,
m.sscmedFcodedvalue, m.Ftype,
m.symptomstatus, m.age_at_eval,
m.measure_Fcode, m.measure_Ftype_revision,
f.multiplex, f.lost_diagnosis, f.sex, f.role, f.Fcode,
f.relationship, f.yearofbirth, f.gender, f.asd, f.scd,
f.language, f.phonological, f.developmental,
f.learning, f.intellectual, f.epilepsy, f.adhd, f.ocd,
f.anxiety, f.depression, f.bipolar,  f.schizophrenia,
f.other, f.age_at_eval, f.measure_Fcode,
f.measure_type_revision, e.multiplex,
e.lost_diagnosis, e.sex, e.role, e.Fcode,
e.school_type, e.grade_level,
e.special_ed_services, e.classroom_setting,
e.personal_aide, e.siblings_intervention_services,
e.siblings_special_ed_services, e.age_at_eval,
e.measure_Fcode, e.measure_type_revision

FROM FollowUpEduTbl as e,
FollowUpFamHisTbl as f, FollowUpMedTbl as
m
WHERE e.sfari_id = f.sfari_id
AND f.sfari_id = m.sfari_id;
```

(d)

Figure 4: SQL Commands: (a) Create Table *FollowUpMedTbl*; (b) Create Table *FollowUpEduTbl*; (c) Create Table *FollowUpFamHisTbl*; and (d) Join operation among all three tables based on sfari_id.

flow "NaturalJoin." After joining three different tables, FollowUpEduTbl, FollowUpFamHisTbl and FollowUpMedTble, we used the final table as our training dataset for our algorithm. Here we also predicted proband's symptom status based on the training set.

Our PR curve for "Symptoms improved" class depicts the prediction accuracy of our algorithm.

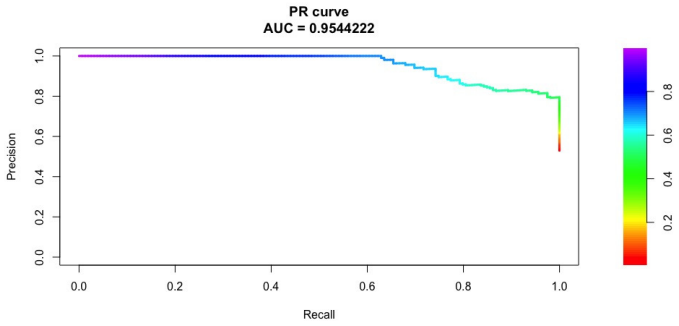

Figure 5: PR curve based on Random Forest Data Mining Technique.

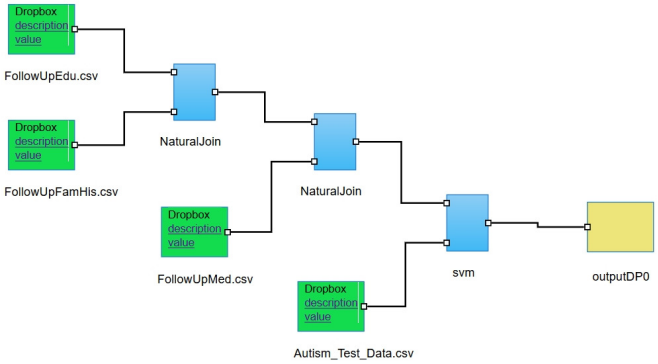

Figure 6: Running Workflow Predicting classes based on Support Vector Machine Data Mining Technique.

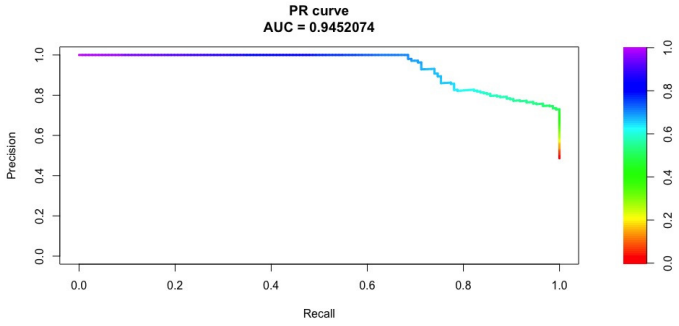

Figure 7: PR based on Support Vector Machine Data Mining Technique.

E. Conditional Random Field Algorithm

We have modeled the problem as a temporal problem with time-stamps associated with symptoms. Thus, we have kept track of patients and how attributes and temporal features have influenced the outcome of the prediction. We observed similar performance on the data set. We did notice a slight improvement in the AUC for "symptoms improved" class. Similar trend is observed for other classes as well. Hence, we postulate that including temporal information captures sequential dependency.

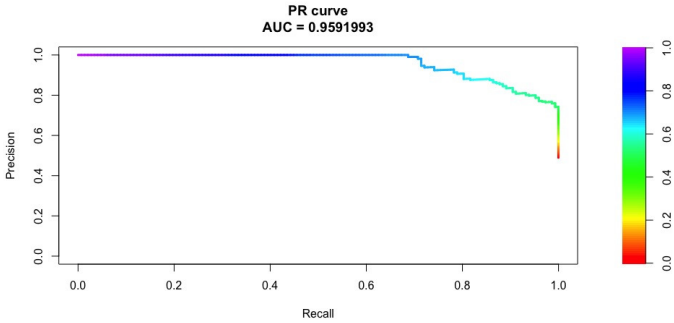

Figure 8: PR curve on conditional random field.

VII. Conclusions and Future Work

In this paper, we present for the first time amalgamation between Autism Health informatics community and Workflow community. This research is motivated by augmenting health informatics into scientific workflows to guarantee data reproducibility. Further research is warranted to capture nuances in the data to guide parents and caregivers in a data-driven manner. To the best of our knowledge, there has not been any study that combines prediction method applied on ASD Phenotype data to guide parents and caregivers.

In the future, we would like to explore our research in the following several directions.

- To identify variables involved and most likely set of variables to have triggered the incident, for each individual episode of manifestation of problem behavior, based on ASD data.
- To recommend next set of goals that are appropriate and beneficial, based on the trend of data.
- To develop tools of built-in construct for facilitating analysis of big data in DATAVIEW platform.

Acknowledgment

This work is supported by National Science Foundation, under grant NSF ACI-1443069. In addition, this material is based upon work supported in part by the National Science Foundation under Grant No. 0910812.

References

[1] T. Latkowski and S. Osowski, "Data mining for feature selection in gene expression autism data," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 864–872, 2015.

[2] R. Guillén, C. Jensen, and S. Edelson, "A machine learning approach for identifying subtypes of autism," in *ACM International Health Informatics Symposium, IHI 2010, Arlington, VA, USA, November 11 - 12, 2010, Proceedings*, 2010, pp. 620–628.

[3] A. Beykikhoshk, O. Arandjelovic, D. Q. Phung, S. Venkatesh, and T. Caelli, "Data-mining twitter and the autism spectrum disorder: A pilot study," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014*, 2014, pp. 349–356.

[4] S. S. Alwakeel, B. Alhalabi, H. M. Aggoune, and M. Alwakeel, "A machine learning based WSN system for autism activity recognition," in *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, 2015, pp. 771–776.

[5] W. Liu, L. Yi, Z. Yu, X. Zou, B. Raj, and M. Li, "Efficient autism spectrum disorder prediction with eye movement: A machine learning framework," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*, 2015, pp. 649–655.

[6] M. P. B. C.-C. L. K. A. Daniel Bone, Matthew S. Goodwin and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, 2014.

[7] M. D. J. Kosmicki, V. Sochat and D. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Translational psychiatry*, vol. 5, no. 2, p. 514, 2015.

[8] J. K. M. Duda and D. Wall, "Testing the accuracy of an observation-based classier for rapid detection of autism risk," *Translational psychiatry*, vol. 4, no. 8, p. 424, 2014.

[9] T. Falck-Ytter and C. von Hofsten, "How special is social looking in asd: a review," *Progress in brain research*, no. 189, pp. 209–222, 2011.

[10] L. Yi, Y. Fan, P. C. Quinn, C. Feng, D. Huang, J. Li, G. Mao, and K. Lee, "Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data," *Journal of Vision*, vol. 13, no. 10, August, 2013.

[11] L. Yi, C. Feng, P. C. Quinn, H. Ding, J. Li, Y. Liu, and K. Lee, "Do individuals with and without autism spectrum disorder scan faces differently? a new multi-method look at an existing controversy," *Autism Research*, vol. 7, no. 1, pp. 72–83, 2014.

[12] H. Lee, T. Hu, H. Jing, Y. Chang, Y. Tsao, Y. Kao, and T. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 215–219.

[13] E. Short, D. Feil-Seifer, and M. J. Mataric, "A comparison of machine learning techniques for modeling human-robot interaction with children with autism," in *Proceedings of the 6th International Conference on Human Robot Interaction, HRI 2011, Lausanne, Switzerland, March 6-9, 2011*, 2011, pp. 251–252.

[14] D. Feil-seifer and M. J. Matari, "Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders," in *In Intl. Symposium on Experimental Robotics*, 2008.

[15] H. Kozima, C. Nakagawa, and Y. Yasuda, "Interactive robots for communication-care: a case-study in autism therapy," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, Aug 2005, pp. 341–346.

[16] B. Robins, K. Dautenhahn, and P. Dickerson, "From isolation to communication: A case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot," in *Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions*, ser. ACHI '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 205–211.

[17] E. Linstead, R. Burns, D. Nguyen, and D. Tyler, "AMP: A platform for managing and mining data in the treatment of autism spectrum disorder," in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2016, Orlando, FL, USA, August 16-20, 2016*, 2016, pp. 2545–2549.

[18] "DATAVIEW: A Big Data Workflow Management System," http://www.dataview.org/.

[19] "Simons Foundation Autism Research Initiative (SFARI)," https://www.sfari.org/.