World Academy of Science, Engineering and Technology
International Journal of Health and Medical Engineering
Vol:17, No:1, 2023

A Machine Learning-based Analysis of Autism Prevalence Rates across US States against Multiple Potential Explanatory Variables

Ronit Chakraborty, Sugata Banerji

Abstract—There has been a marked increase in the reported prevalence of Autism Spectrum Disorder (ASD) among children in the US over the past two decades. This research has analyzed the growth in state-level ASD prevalence against 45 different potentially explanatory factors including socio-economic, demographic, healthcare, public policy and political factors. The goal was to understand if these factors have adequate predictive power in modeling the differential growth in ASD prevalence across various states, and, if they do, which factors are the most influential. The key findings of this study include (1) there is a confirmation that the chosen feature set has considerable power in predicting the growth in ASD prevalence, (2) the most influential predictive factors are identified, (3) given the nature of the most influential predictive variables, an indication that a considerable portion of the reported ASD prevalence differentials across states could be attributable to over and under diagnosis, and (4) Florida is identified as a key outlier state pointing to a potential under-diagnosis of ASD.

Keywords—Autism Spectrum Disorder, ASD, clustering, Machine Learning, predictive modeling.

I. Introduction

ASD is a developmental disability that may cause affected individuals to behave, communicate and learn in ways that are different from others. ASD is known to be caused by differences in the brain during development, but the reason for the occurrence of these differences are not all known. According to the Centers for Disease Control and Prevention (CDC)'s Autism and Developmental Disabilities Monitoring (ADDM) network, 1 in 44 children were identified to have ASD in 2018 among a sample of 8 year old children across the United States. There has been a clear upward trend in ASD prevalence numbers among children in the US across geographical sites over the past two decades as shown in Fig. 1. These numbers are seen across multiple data sources and hence cannot be attributed to data collection artifacts. It is, however, not apparent how much of this increase is attributable to changes in the clinical definition of ASD, better diagnostics efforts or actual increases of prevalence.

In this study, an attempt was made to understand the ASD trend across states in the US and 45 different potential factors that might have a correlative and/or causative relationship with ASD prevalence were analyzed. Both unsupervised and supervised machine learning techniques were used to determine the degree of predictive power that these explanatory factors have when it comes to ASD prevalence and some discussion was provided regarding the presence of potentially causative relationships where a predictive link was observed.

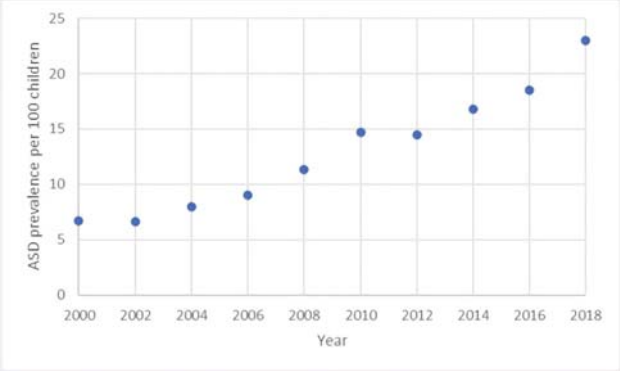

Fig. 1 ASD Prevalence estimates (in terms of 1 in 1000 children) in the US over time [1]

II. Previous Work

There have been several research studies over the years that have tried to establish correlation and causal links between multiple explanatory variables and the rise in reported ASD prevalence. Reference [2] investigated genetic and environmental factors (the effect of exposure to neurotoxins such as mercury and lead) during critical stages in a child's early development. It applied Combinatorial Fusion Analysis (CFA) and Association Rule Mining (ARM) to ASD prevalence, mercury, and lead data to identify potential associations. Research [3]–[5] explored potential association between genetic factors and ASD prevalence and identified genes that may be linked to ASD. Reference [6] found indications that certain genetic vulnerabilities, such as reduced ability to excrete mercury, and exposure to mercury during a child's development, may cause higher incidences of ASD. Many recent studies [7]–[11] using statistical analysis, clustering and machine learning techniques have focused on applying these in better diagnosing ASD from EI (early intervention) records, ERG (electroretinogram), MRI (magnetic resonance imaging), f-MRI (functional magnetic resonance imaging), electrocardiogram (ECG), skin

Ronit Chakraborty worked as a research intern at Lake Forest College, Lake Forest, IL 60045, USA (corresponding author, e-mail: ronit.chakraborty05@gmail.com).

Sugata Banerji is an associate professor at the Department of Mathematics and Computer Science, Lake Forest College, Lake Forest, IL 60045, USA (e-mail: banerji@lakeforest.edu).

conductance (SC), respiration and skin temperature, pre-verbal vocalizations, and ammonia concentrations. However, the proposed study is unique in (1) applying the clustering and machine learning techniques to help identify the underlying factors that may be associated with differential ASD prevalence rates observed across US states, and (2) identifying the most influential factors for predicting overall ASD prevalence.

III. Methodology

The goal of the proposed study was to analyze the explanatory power of the candidate independent/explanatory factors in predicting the ASD prevalence rate in a given US state. This problem was approached from a few different angles. Since the factors data had high dimensionality (45 different factors), t-distributed stochastic neighbor embedding (t-SNE) [12], [13], a powerful dimensionality reduction technique, was utilized to map the independent variables to a 2-dimesional space and analyze for any clustering of the states. If the states tend to form clusters, and the clusters tend to have similar ASD rates, that would indicate a strong predictive relationship between the factors and the ASD rate.

The direct correlation coefficient between each factor and the ASD rate was computed to evaluate factors that may have strong positive or negative correlation with ASD prevalence.

Finally, the data from different states were split into random training and test sets, and random forest regressors [14], a flexible machine learning technique, was leveraged to fit a model to the training data and then to predict ASD prevalence from the test set. A high efficacy of the model would indicate strong predictive relationship between the factors and the ASD prevalence.

The study selected a wide variety of factors/potentially explanatory variables from a number of areas:

1) Socio-economic: median income, population with income at various multiples of poverty level
2) Demographic: total population, breakdown by ethnicity
3) Healthcare: availability of physicians including psychiatrists, hospital beds of various types, insurance coverage, infant mortality, newborns with low birthweight, pre-term births, ease/difficulty of accessing mental health services, prevalence of mental illness/depressive episodes
4) Public policy: Medicaid spending, public spending on various levels of education and healthcare
5) Political: voter registrations, voting percentages

IV. Data Sources

ASD prevalence numbers for 2000-2018 were obtained for the network of sites across 17 states monitored by the CDC ADDM network [1].

References [15], [16] were used to source the data on birth with low birth weight per state, children who received mental healthcare per state, children who had difficulty getting mental care, areas with shortage of mental health care professionals, children who needed mental care but did not receive it, children whose parents had difficulty paying medical bills,

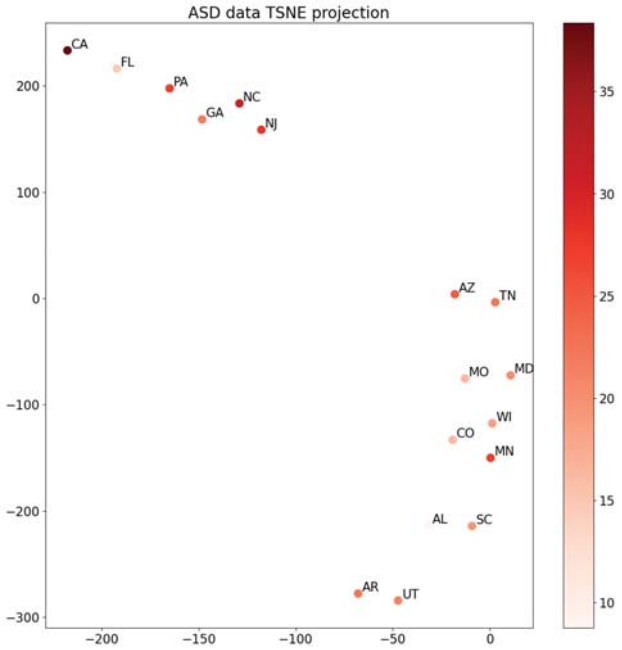

Fig. 2 Projection of 45 Factors into 2 Dimensions Using t-SNE

children who had both medical and dental preventive care visit in 12 months, children by health insurance status (employer, Medicaid, other, uninsured), per capita healthcare expenditure by service (hospital, physician, dental, drugs, other), hospital admissions by bed type (state/public, nonprofit, for-profit), hospital beds by bed type (state/public, nonprofit, for-profit), infant mortality, Medicaid CHIP eligibility as a percent of federal poverty level, Medicaid/CHIP participation rate among eligible children, total Medicaid spending, children who received any mental health care in 12 months, percent of children by race, preterm births as a percent of all births, total expenditure by area (elementary/secondary education, higher education, public assistance, Medicaid, corrections, transportation), adults reporting mental illness (any and serious) in 12 months, adults reporting unmet mental care need, adolescents reporting depressive episodes, number of active specialists (including psychiatrists), number of active physicians (primary care and specialist), percentage of population at 100%, 200%, and 400% of poverty level, and percentage of voter population that registered and voted.

From [17], [18] the state-wise population, racial makeup and income data was obtained. A complete list of all the factors considered for this study can be found in Table I.

V. Analysis Approach and Results

A. Dimensionality Reduction and Clustering

Given the high dimensionality (45 dimensions) of this factors data set, a decision was made to utilize an efficient dimensionality reduction technique that can map the high number of dimensions into 2 or 3 dimensions while preserving the similarity distance among the data points such that similar objects remain close to each other and dissimilar objects

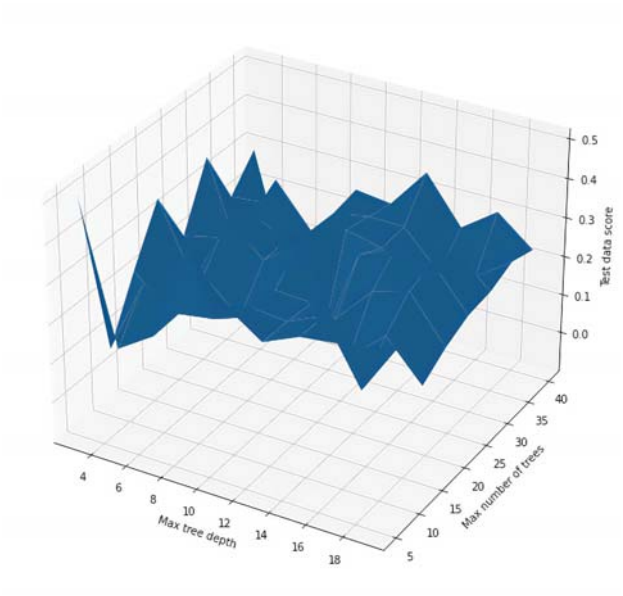

Fig. 3 Predictive score of the Random Forest models on the test data set for various parameter values

remain at a distance with a high probability. T-distributed stochastic neighbor embedding (t-SNE) [12] [13], a method that maps high-dimensional data into a 2 or 3-dimensional map for easy visualization, was found to be an appropriate method for this analysis. After the dimensionality reduction, it becomes easier to verify if the data points form any clusters which would indicate the potential for classification of data points based on the factors. Python's scikit-learn package [19] was used to implement the t-SNE dimensionality reduction and clustering analysis.

The results of the t-SNE projection of the ASD prevalence data set and mapping 45 potential explanatory factors into 2 dimensions is shown in Fig. 2. Each state is marked with a color that is drawn from a spectrum of red where deeper red indicates high ASD prevalence and lighter red indicates low ASD prevalence. Interestingly, the states form two distinct clusters, one made up of states with generally higher prevalence of ASD at the top left corner of the graph, and the other containing states with generally lower prevalence of ASD at the bottom right corner of the graph. The cluster at the top left of the figure represents 6 states, 4 of which have a ASD prevalence above 25 (per 1000). Only Florida has a prevalence below 20 in the cluster. The cluster at the bottom right of the figure includes 11 states all of which, except Minnesota, have ASD prevalence below 25. This clustering provides a strong indication that the factors under study have a substantial predictive value for ASD prevalence.

B. Correlation Statistics

The correlation coefficient of each factor against ASD prevalence rates was also computed to understand any positive or negative correlation that may exist. The results are captured in Table I. Four factors demonstrate a positive correlation above 0.5 and another four demonstrate a negative correlation less than -0.5. The factors with highest positive correlation are: percentage of children of Asian/Native Hawaiian/Pacific Islander ethnicity in a state, Medicaid spend per capita, per capita spending on corrections, and state population. The factors with the highest negative correlation are: hospital admissions to for-profit beds, percentage of children without insurance, percentage of births with low birth weight, and percentage of births that are pre-term.

TABLE I
CORRELATION OF DIFFERENT POTENTIAL FACTORS AGAINST ASD PREVALENCE

Factor	Correlation
% of children who are Asian/Native Hawaiian and Pacific Islander	0.71
Medicaid spending per capita	0.57
Corrections spending per capita	0.57
Population	0.54
Elementary secondary education spending per capita	0.47
% of children who are Hispanic	0.44
Medicaid spending per capita (excluding adm costs)	0.41
% of children of multiple races	0.40
Median household income	0.38
Psychiatrists per million population	0.36
% of children who received mental care	0.35
Hospital admissions per 1000 (non-profit)	0.30
% of voter population who voted	0.27
% of families with income at 400% of poverty level	0.24
Hospital beds per 1000 population (non-profit)	0.23
% of voter population registered	0.23
Physicians per million population	0.22
% of children who are American Indian/Alaska Native	0.16
Health spending per capita	0.14
Healthcare expenditure per capita	0.14
Medicaid CHIP child participation rate	0.11
% of children with insurance from employer	0.06
% of children with Medicaid insurance	0.02
% of children with difficulty accessing mental care	-0.04
% of Adults reporting any mental illness	-0.06
% of adolescents with major depressive episodes	-0.07
% of children with insurance (non-group)	-0.09
% of children with medical/dental preventive care visits in last 12 months	-0.10
% of families with income at 100-199% of poverty level	-0.14
% of families with income under poverty level	-0.21
% of families with income at 200-399% of poverty level	-0.28
Higher education spending per capita	-0.29
Hospital admissions per 1000 population (total)	-0.32
% of children who are Black	-0.32
Hospital beds per 1000 population (total)	-0.36
% of children who are White	-0.36
Psychiatrist shortfall%	-0.44
% of children whose parents had difficulty paying child's medical bill	-0.44
Hospital beds per 1000 population (state/local government)	-0.46
Hospital beds per 1000 population (for-profit)	-0.47
Infant mortality rate per 1000	-0.48
% of preterm births	-0.55
Low birth weight %	-0.55
% of children who are uninsured	-0.57
Hospital admissions per 1000 (for-profit)	-0.58

C. Model Fitting Using Random Forest Regressor

Random forests [14], a flexible machine learning model, were evaluated on the ASD data set. To train the model, 70% of the data were randomly selected and the other 30% were used to test the model. Random forests are an ensemble machine learning technique for both classification and regression problems. It works by constructing a multitude of decision trees on the training data and then generating the output based on the class selected by the most number of trees (for a classification problem), or the average of the values returned by the trees (for a regression problem). The use of many trees in a forest helps avoid the tendency of single decision tree models to overfit to the training data. The random forest regressor was chosen for this modeling problem

as they are particularly well-suited to closely approximate any arbitrarily complex n-dimensional regression function and does not require elaborate tuning of the hyper parameters. An added advantage of the random forests is the fact that because the final prediction is based on the output of individual decision tress, the results are highly interpretable, and provide useful insight and intuition into the nature of the data, including the relative importance of the features.

A model's efficacy is measured using the prediction score [20], γ, in (1):

$$\gamma = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} \quad (1)$$

where y_i represents the actual value of the dependent variable for the i^{th} data point, $\hat{y}_i$ represents the predicted value of the i^{th} data point by the model, $\bar{y}$ represents the mean value of the dependent variable, and $n+1$ is the total number of data points. γ measures how much better the chosen model is in predicting the dependent variable compared to a brute-force approach of using the mean value of the dependent variable as the prediction.

Given the limited depth with high width of the data set, many iterations were run over the random forest parameter space with varying limits on the maximum depth of each tree as well as the total number of trees. The resulting predictive scores of the models on the test data have been plotted on a 3-dimensional surface graph in Fig. 3.

As seen from the surface plot, the highest predictive score is achieved on the test dataset at maximum tree depth = 3 and maximum number of trees = 5. The highest predictive efficacy obtained is 51%.

The *feature_importances* attribute of the random forest regressor was used to determine the relative importance of different explanatory variables. The relative importance of top 5 features are shown in Table II.

D. Creating a Parsimonious Model

In order to evaluate if an equivalent predictive performance can be obtained while using a smaller set of factors, first the cross-correlations between each pair of independent variables was computed. Fig. 4 shows a heatmap of the pair-wise cross-correlation of the factors in the dataset. The pairs that have a high positive (> 0.9) or negative (< -0.9) correlation can potentially be replaced by just one variable from the pair. The pairs with the highest positive or negative correlations are shown in Table III.

TABLE II
RELATIVE IMPORTANCE OF TOP 5 FEATURES BASED ON THE RANDOM FOREST REGRESSOR

Feature	Relative Importance
% of children with difficulty accessing mental care	0.21
% of children receiving mental care	0.16
Medicaid spending per capita	0.15
% of children who are Hispanic	0.14
% of children who are uninsured	0.12

TABLE III
FACTOR-PAIRS WITH HIGHEST MAGNITUDE POSITIVE OR NEGATIVE CROSS-CORRELATION

Factor 1	Factor 2	Cross-correlation
Healthcare expenditure per capita	Health spending per capita	1.000
Median household income	Income 400% of poverty level	0.904
Hospital beds per 1000 (total)	Hospital admissions per 1000 (total)	0.918
Hospital beds per 1000 (non-profit)	Hospital admissions per 1000 (non-profit)	0.964
% of voter population voted	% of voter population registered	0.964
% of Children with insurance(employer)	% of children with Medicaid/insurance	-0.937
% of people at 100-199% of poverty level	% of people above 400% of poverty level	-0.930

As 7 factor pairs were found above the correlation threshold, one factor from each of those pairs were dropped. It may be noted that one of these pairs is perfectly correlated (with a correlation coefficient of 1.0). This is because these two factors, in spite of being collected from two different sources, were found to contain identical data. This reduced the number of factors from 45 to 38.

With this smaller set of factors (38 factors), the random forest regressor model demonstrated a highest efficacy of just above 50% on the test dataset. This is nearly equivalent to the highest efficacy obtained with the original set of 45 factors. So, this study is able to confirm that a shortened set of explanatory factors may be utilized for predicting ASD prevalence without losing model efficacy.

VI. DISCUSSION AND FUTURE WORK

Clustering analysis using the potential factors shows clear bunching of states into two groups with highest and lower ASD prevalence rates. This indicates a strong link between the factors and the ASD prevalence. The state that seems a prominent outlier is Florida. It has a relatively low (14.45 in every 1000 children) reported ASD prevalence whereas all the explanatory factors associated with that state make it a member of the higher ASD prevalence cluster. This potentially points to an under-diagnosis of ASD in the state of Florida relative to its peers.

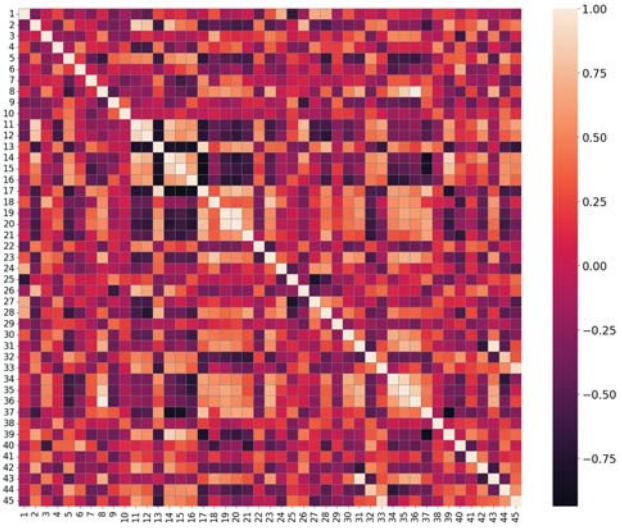

Fig. 4 Heatmap of cross-correlations between each pair of factors

The correlation analysis indicates that the factors with highest (> 0.5) positive correlation to ASD prevalence are: (1) percentage of children of Asian/Native Hawaiian/Pacific Islander ethnicity in a state, (2) Medicaid spend per capita, (3) per capita spending on corrections, and (4) state population. Out of these, only the second can be thought of having a direct impact on a state's health metrics, and one would expect that impact to be positive. The factors with the highest magnitude (< -0.5) negative correlation to ASD prevalence are: (1) hospital admissions to for-profit beds, (2) percentage of children without insurance, (3) percentage of births with low birth weight, and (4) percentage of births that are pre-term. Out of these, the last three can be thought of having a direct impact on a state's health metrics, and one would expect that impact to be negative. However, the observed high positive and negative correlations appear to be working in the exact opposite direction of their expected impact. In other words, states with better underlying health related factors are reporting higher ASD prevalence rates and vice-versa. This leads us to believe that ASD is under-diagnosed in states with weaker health related factors compared to states with otherwise stronger health metrics.

The machine learning based analysis using random forests indicates a relatively high (51%) efficacy of the model. This demonstrates that the chosen explanatory factors have a high degree of efficacy in predictive ASD prevalence rates.

As part of future research, a study is planned to analyze temporal evolution of the explanatory variables and the ASD prevalence over several years to understand if there are any discernible trends in the dynamics. An expansion of the scope of this study is also planned to include additional states beyond the 17 that are covered by CDC's ADDM network by potentially using other sources of data to estimate state-level ASD prevalence rates.

REFERENCES

[1] "Autism and Developmental Disabilities Monitoring (ADDM) - Centers for Disease Control and Prevention," https://www.cdc.gov/, accessed on Fri, November 11, 2022. [Online]. Available: https://www.cdc.gov/ncbddd/autism/data/index.html

[2] C. Schweikert, Y. Li, D. Dayya, D. Yens, M. Torrents, and D. Hsu, "Analysis of autism prevalence and neurotoxins using combinatorial fusion and association rule mining," 06 2009, pp. 400–404.

[3] R. Nataf, C. Skorupka, L. Amet, A. Lam, A. Springbett, and R. Lathe, "Porphyrinuria in childhood autistic disorder: Implications for environmental toxicity," Toxicology and Applied Pharmacology, vol. 214, no. 2, 2006, pp. 99–108.

[4] R. A. Kumar, S. KaraMohamed, J. Sudi, D. F. Conrad, C. Brune, J. A. Badner, T. C. Gilliam, N. J. Nowak, J. Cook, Edwin H., W. B. Dobyns, and S. L. Christian, "Recurrent 16p11.2 microdeletions in autism," Human Molecular Genetics, vol. 17, no. 4, 12 2007, pp. 628–638.

[5] L. A. Weiss, Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B.-L. Wu, and M. J. Daly, "Association between microdeletion and microduplication at 16p11.2 and autism," New England Journal of Medicine, vol. 358, no. 7, 2008, pp. 667–675.

[6] D. A. Geier, P. G. King, L. K. Sykes, and M. R. Geier, "A comprehensive review of mercury provoked autism." The Indian journal of medical research, vol. 128 4, 2008, pp. 383–411.

[7] M. Liu, Y. An, X. Hu, D. Langer, C. Newschaffer, and L. Shea, "An evaluation of identification of suspected autism spectrum disorder (asd) cases in early intervention (ei) records," 2013, pp. 566–571.

[8] S. M. Manjur, M.-B. Hossain, P. A. Constable, D. A. Thompson, F. Marmolejo-Ramos, I. O. Lee, D. H. Skuse, and H. F. Posada-Quintero, "Detecting autism spectrum disorder using spectral analysis of electroretinogram and machine learning: Preliminary results," in 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 3435–3438.

[9] B. S. Roopa and R. Manjunatha Prasad, "Concatenating framework in asd analysis towards research progress," in 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, pp. 269–271.

[10] J. F. Santos, N. Brosh, T. H. Falk, L. Zwaigenbaum, S. E. Bryson, W. Roberts, I. M. Smith, P. Szatmari, and J. A. Brian, "Very early detection of autism spectrum disorders based on acoustic analysis of pre-verbal vocalizations of 18-month old toddlers," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7567–7571.

[11] M. B. Marchelliant, Aripin, and S. A. Wulandari, "Analysis of electrocardiogram signal and ammonia concentration for clustering asd condition," in 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), 2021, pp. 290–295.

[12] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in Advances in Neural Information Processing Systems, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2002.

[13] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research, vol. 9, no. 86, 2008, pp. 2579–2605. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[14] T. K. Ho, "Random decision forests," in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, 1995, pp. 278–282 vol.1.

[15] "Kaiser Family Foundation: Hospital Beds by State ," https://www.kff.org, accessed on Fri, November 11, 2022. [Online]. Available: https://www.kff.org/other/state-indicator/beds-by-ownership

[16] "Kaiser Family Foundation: State health data ," https://www.kff.org, accessed on Fri, November 11, 2022. [Online]. Available: https://www.kff.org/statedata/

[17] "US Census Bureau: Population and housing unit estimates," https://www.census.gov, accessed on Fri, November 11, 2022. [Online]. Available: https://www.census.gov/programs-surveys/popest/data/data-sets.html

[18] "Household income data by state and race," https://www.census.gov, accessed on Fri, November 11, 2022. [Online]. Available: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

[19] "scikit-learn Machine Learning library in Python," http://scikit-learn.org, accessed on Fri, November 11, 2022. [Online]. Available: http://scikit-learn.org/

[20] S. V. Chakraborty and S. K. Shukla, "Predictive modeling of electricity trading prices and the impact of increasing solar energy penetration," in 2019 IEEE Milan PowerTech, 2019, pp. 1–6.