

# An Investigation on Machine Intelligent Techniques to Regularize Incompetency in Communiqué

*S. Thirumal, Research Scholar, Hindustan Institute of Technology and Science, Chennai.*

*Assistant Professor, Department of CSE, VISTAS.*

*J. Thangakumar, Associate Professor, Department of CSE, Hindustan Institute of Technology, Chennai.*

*D. Venkata Subramanian, Adjunct Professor, Department of CSE, Velammal Institute of Technology, Chennai.*

**Abstract---** The analytics of Big Data in the field of healthcare and medicine has enabled large datasets to be analysed from several patients thus identifying the correlation among datasets and their clusters resulting in the development of predictive models with techniques of data mining. Communication disability has been associated with issues and disability of learning. Almost 70% of those with a condition of communication disability seem to have a certain level of learning disability that is based on the IQ score of the individual and also their cognitive functioning levels. Another genomic birth related syndrome known as the Down syndrome is the result of either all or a portion of an extra copy of the chromosome 21 which continues during the entire lifetime of a person. In this work, a survey of works available in the literature that are based on Down syndrome and autism affected persons by employing techniques of Big data or Machine learning is reviewed.

**Keywords---** Big Data Analytics, Autism, Communication Disability, Down Syndrome.

## I. Introduction

The ecosystem of Big Data is a compendium of all the elements covering various aspects of management of Big Data and its analytics. It permits one to be able to move, to store, analyse, process, visualize and also manage data. The technologies of Big Data have broken the conventions that were established earlier and have also helped all industries to completely uncover older patterns of information, proposing questions and finally deliver a service that is decent.

A condition of Autism will normally occur in about 1-2% of the population and the wellbeing of individuals with autism has become a primary issue in public health. It is a diagnostic label which is based on human behaviour. As the criteria attempts at maximizing some clinical consensus, it also is able to mask a large level of heterogeneity among individuals in different analytical levels.

Down Syndrome (DS) on the other hand, is a genetic disease which is connected to chromosomal abnormality (a trisomy in the Hsa21/chromosome 21 in the human body). This often results in a disability of intellect and affects about 1 in a total of 1000 live births that take place in the world. The Hsa21 further encodes above 500 genes which are an unknown subset that may be connected to issues in learning. Meantime which is the drug registered for the treatment of the Alzheimer's Disease (AD) was able to provide potential to manage learning deficit in cases of DS. There was also a new mouse protein expression dataset that was available consisting of 77 proteins that were measured in the cerebral cortex of a total of 8 classes to control the Down syndrome mice (Ts65Dn) that were exposed to a condition known as the Context Fear Conditioning (CFC). For this, there was a task employed for assessing associated learning either with or sometimes without treatment using mamefine. Some studies conducted recently proved the manner in which the dataset may be used for clustering such proteins that were based on any different in the levels of protein found in different types of conditioning [13].

## II. Literature Survey

Cai et al [1] have taken Big Data in healthcare to be the object of research. The source for health care in Big Data was introduced to summarize the current status of the application and further analysed the different challenges faced and the opportunities available in medical Big Data. Developing of Big Data has several other challenges in various aspects such as platform development, training of personnel, the security of data and data cleaning.

Recently, an observation has been made with regard to a move that was encouraged by stakeholders to develop open data in various research areas. One of the most crucial areas where the Big open data can be quite useful was heterogeneous disorders which were complex like the Autism Spectrum Disorder (ASD). There were several inconsistencies in findings and the high heterogeneity of the ASD had to make use of the Big and the open data for handling crucial challenges like defining and understanding the heterogeneity and also the other subtypes of the ASD. For this, there were several initiatives taken aiming at developing Big open data for research in autism. For the purpose of providing a useful reference of data for these researchers, there was a systematic search made for the data resources of the ASD by employing Scopus database, and Google search engine along with pages of 'recommended repositories' from key journals. There was a review for making a systematic search of ASD data of several types of data by Al-jawahiri and Milne [2]. The types of data were, the phenotypic, the neuro imaging, connectivity matrices of the human brain, statistical maps of the human brain, recruitment of ASD participants and bio-specimens. There were about 33 resources with various types of data with different participants observed.

Describing the available data from each of the resources and providing links to every such resource was made. Furthermore, the primary implications were addressed and data that was under presented were identified. A genetic disorder that is the result of trisomy of a part of or the entire human chromosome 21 is termed as Down syndrome (DS). As there is no remedy for the condition of Down Syndrome for which screening tests were found to be extremely efficient in preventing a condition of DS. For this, Feng et al [3] had made use of learning algorithms that were supervised for building the DS classification that was based on either the protein or the protein modification and expression level of the mice of DS model Ts65Dn. Also, there was another method known as the adaptive boosted Decision Tree that was applied for identifying the informative and correlated protein factors connected to the biological process of the DS and their pathways. Furthermore, the DS classification was improved and their screening models by means of using the chosen DS and the critical proteins that were connected. Lastly, learning algorithms that were unsupervised were employed for confirming results that were obtained. The chosen DS and related proteins were used further for the gene-related drugs and their development.

Pharmacotherapies for the intellectual disability (ID) were generally not known as the different abnormalities at a complex and molecular level causing the ID that were challenging to be able to understand. The Down syndrome (DS) is the primary cause of ID and the reason for this was the extra copy of a human chromosome 21 (Hsa21) investigated on the levels of protein by means of employing the Ts65Dn based mouse model of the DS that were the orthologs of about 50% percent of the Hsa21 with the protein-coding genes. In all recent studies, methods of classification were applied for understanding the critical aspects of DS factors. There was a method of forward feature selection that was applied to identify all correlated proteins and also their DS interaction that was introduced by Kulan and Dag [4]. Once the identification is complete, the supervised learning model for the levels of expression for understanding all critical proteins to diagnose and further explain the DS. The technique proposed had depicted all results of optimum classification by means of parameters of optimization within the grid search. On being compared to the earlier work, the results of classification had a higher level of accuracy.

Koivu et al [5] had made an evaluation of the algorithms of machine learning for improving the Down Syndrome screening in the first trimester. There are two different real-world datasets that were employed for being used in the experiment along with multiple algorithms for classification. The models that were implemented had been tested in a third and real-world dataset after which it was compared to the predicate method which was software for risk assessment. The Deep Neural Network based model provided a curve which was about 0.96 and with a rate of detection of about 78% that was given under a curve of about 0.95 and a rate of detection of 61% with about 1% of the false positive rate. On being compared to this predicate method, the best model of support vector was found to be inferior and the optimized neural network based model could provide a higher rate of detection along with a false positive rate with a similar rate of detection and a lower false positive rate. The findings were able to improve the screening of the first trimester of the Down syndrome using certain clinical variables along with the huge training data that was obtained from a particular population.

Research on Autism spectrum disorder (ASD) needs to leverage Bbig data^ on a scale similar to the other fields. There has been a significant increase in the evaluation of research literature that evaluates machine learning effectiveness in the diagnosis of ASD. There was a comprehensive review made on a total of 45 papers that made use of the ASD based supervised machine learning methods that included text analysis by Hyde et al [6]. The primary goal of this paper was the identification and the description of the trends in supervised machine learning for guiding researchers that are interested in the expansion of the body of the computationally, clinically and statistically strong approaches in the mining of ASD data.

Ramanathan et al [7] had focused on the extraction of markers from Prenatal Screening (the test of the first trimester) reports of these patients. Some of these attributes were smoking habits, age, nuchal translucency, length of crown-rump and history of earlier pregnancies along with Trisomy 21 with the presence of a nasal bone extracted by employing Optical Character Recognition. Collection of these reports were subject to different techniques of clustering such as the K-means, the K-medoids, Hierarchical clustering and the DBSCAN (Density-based spatial clustering of applications with noise). The paper is targeted towards a technique to handle the skewed datasets by employing the ADASYN (Adaptive Synthetic Sampling Approach) based approach of over-sampling in order to handle class bias owing to the dataset that rarely has 'High Risk' DS. These reports will get subject to the ML (Machine Learning) based ensemble to supervised learning methods such as the ANN (Artificial Neural Network), Random Forest and Naive Bayes for the determination of the posterior probability of a foetus that may suffer from a condition of Down syndrome. This paper further envisions the building of a new architecture which may be used as a new tool that helps the gynaecologists to provide a measure of the numerical probability that represents the risk of Down Syndrome in a foetus. A gynaecologist may recommend if a patient needs to take up some invasive [amniocentesis with the CVS (Chorionic Villus Sampling)] based tests that will be able to stand as a clear indicator of the DS. This paper makes use of live datasets and can reflect the DS scenario in India. It aims at providing path-breaking ideas in prenatal diagnosis that makes use of the methodologies of ML.

Çelik et al [8] had detected and further estimated this Down Syndrome by means of an analysis of the levels of protein in the genes. This way, there was a new system of Decision support that was based on the techniques of machine learning that were proposed for the estimation of conditions of Down Syndrome. In addition to this, there was yet another technique called the Principal Component Analyses that were performed for the elimination of multi-proteins found in genes in a lower number for achieving similar success with limited information.

Furat and Ibriki [9] brought about samples that had Down Syndrome and did not have it in the dataset of mice protein expression. In the current study, this dataset from the UCI repository was classified on the basis of the Support Vector Machine, Random Forest, Decision Table, K-Nearest Neighbour, and the Bayesian Network Algorithm. These algorithms of classification that had 10fold cross validation were split equally in the form of test data and train data for classifying the mice protein dataset.

This data classification was successful with an accuracy of about 94.3519% in 0.06 seconds by employing the Bayesian Network, and accuracy of 99.2593% in about 0.01 seconds with the KNN, an accuracy of 95.4630 % in about 1.2 seconds with the Decision Table, an accuracy of 100% in about 0.58 seconds with the Random forest and 100% accuracy in about 1.17 seconds with the SVM and 10-fold cross-validation. At the same time, the dataset classification was successful with an accuracy of about 95.3704% in 0.22 seconds by using the Bayesian Network, with an accuracy of about 98.3333% in 0 seconds by using the KNN, with an accuracy of about 98.3333% accuracy in about 0.72 seconds by using the Decision Table, with an accuracy of 100% in 0.77 seconds by using the Random Forest and finally an accuracy of 100% accuracy in about 1.48 seconds by using the SVM, by means of using the train-test data division equally.

The primary goal of this research was to be able to example the capacity of analytics of Big Data in genetic studies. The Cloudera's Hive which was the core component obtained from the Hadoop ecosystem for the purpose of uploading, applying schema and processing it was introduced by Feng [10]. On the basis of the protein dataset expression profile, a common key molecule may be identified which was in response to the memantine which was the marketed drug for a condition of Alzheimer's Disease, used in the mouse model of Down syndrome. Such findings help in providing newer strategies for validation by Data Scientists all experimental data efficiently.

Lombardo et al [11] had made a new presentation of principle organization that examined autism in multi-level heterogeneity. All theoretical concepts like 'autisms' and 'spectrum' will reflect only the explanations that are non-mutually exclusive with regard to the dimensional and continuous variation among individuals. But there are some common practices involving studies of smaller samples along with models of case-control that are suboptimal to tackle heterogeneity. Big Data has become a critical ingredient to furthering any more understanding of autism heterogeneity. It is also feature-rich and broad aside from deep as well. Such characteristics are helpful in ensuring results that can be generalizable and can also facilitate utility evaluation. The utility of the model is shown by the ability it has to explain in a clinical manner all important phenomena. It has a directionality to explain any variability across various levels which may either be top-down or bottom-up. There is progress observed with these supervised models that have been built on a priori and sometimes predicted theoretically as it can become crucially important to be able to complement this kind of work along with discoveries which are driven by data were leverage is unknown having multivariate distinctions in Big data. If the manner in which heterogeneity is modelled among autistic persons is not understood, there may be only limited progress in the precision medicine goal.

The Autism Spectrum Disorder (ASD) has been characterised by phenotype heterogeneity that is viewed to be an obstacle to studying etiology, treatment, prognosis, and diagnosis. The concept of heterogeneity in the ASD was both complex and multidimensional and it also includes variability in the phenotype, pathologic, physiologic, and clinical parameters. Obafemi-Ajayi et al [12] had applied a model of hierarchical clustering that was well-suited to deal with all types of datasets that stratify children having ASD into subgroups that are homogeneous that were in line with Diagnostic, as well as Statistical Manual of Mental Disorders (DSM)-5 model. These results provided a better level of understanding of various complex issues with ASD phenotypic heterogeneity. The goal here was to provide an insight into all viable genotypic and phenotypic markers to guide cluster analysis for the ASD genetic data. They further analysed clusters with their hierarchical structure and were suitable as a model that could unravel this disorder and its complex heterogeneity.

For the purpose of tackling heterogeneity among children having autism, there are recent advances found in deep learning that have been employed for formulation of a framework of personalized machine learning (ML) used for an automatic perception of the affective state of children at the time of engaging them in autism therapy assisted by robots. As opposed to making use of the concept of one-size-fits-all, there was a personalized framework for every child based on contextual information (the demographics and the scores of behavioural assessment) and their individual characters proposed by Rudovic et al [13]. Evaluating this framework based on a multimodal video or audio and sometimes autonomic physiology using a dataset consisting of 35 children falling between the ages of 3 and 13 having autism were considered. These were obtained from two different cultures which were Europe and Asia that achieved average agreement among intra-class correlation of about ~60% having human experts to estimate and further affect engagement thus outperforming all ML solutions that were non-personalized. Their results demonstrated the robot perception feasibility that affects the engagement of children who are autistic with implications for designing therapies for autism in the future.

The techniques of Active Machine Learning (AML) can enable a model of machine learning that performs well using by training data. Another new AML approach to teaching the concept of object recognition among children with ASD and the effects were compared to Passive Learning (PL). There was another application known as the web and tough application that was developed for the purpose of teaching object recognition in cases where the objects had been grouped in accordance with their levels of difficulty. The process of teaching had been based on the principles of Applied Behavioural Analysis where five children with a condition of mild or moderate ASD were taken as participants. There was a design of alternating treatments for research of single subjects that was employed. Results showed that the AML had been very effective compared to the PL for at least four of five participants. As a consequence, they are able to learn at a faster pace with lesser trials needed for reaching the criterion of learning.

Using machine learning techniques when there is the absence of expertise of the clinical domain, may become tenuous and result in conclusions that are misinformed. For illustrating this, Bone et al [15] had made a critical evaluation of reproduce results from two different studies (Wall et al. in *Transl Psychiatry* 2(4): e100, 2012a; *PloS One* 7(8), 2012b) which claimed a drastic reduction in time for diagnosis of optimism with machine learning.

If there is some failure in generating findings that are comparable to the ones that were reported by Wall and his colleagues along with methodological and conceptual issues connected to this study. To conclude, using the proposed practices while using techniques of machine learning in the research of autism, some of the areas that are promising for collaborative work in the intersection of behavioural science and computational science.

Even though the incidence of Autism Spectrum Disorder (ASD) with Attention Deficit Hyperactivity Disorder (ADHD) has been on the rise affecting about 410% of the population of paediatrics, its diagnosis continues to remain subjective, time-intensive and cumbersome. Since there are upward gaps of a year between the suspicion and its actual diagnosis, time for treatment and interventions is lost. Therefore, there is a need to make quick assessments of risk to streamline this process. employing methods of feedforward selection and under-sampling have helped in training six models of machine learning by Duda et al [16] that was tested and trained on the entire 65-item Social Responsiveness Scale based score sheets from a total of 2925 individuals having ASD (n = 2775) or ADHD (n = 150). It was identified that a total of 5 among 65 of the behaviours that were measured by means of the screening tool was enough for distinguishing the ASD from the ADHD with a high level of accuracy (area under the curve of 0.965). The results further support hypothesis as follows: (1) Machine Learning may be used as a discern between ASD and ADHD with a high level of accuracy (2) such distinction is made with a small set of behaviours measured commonly. Findings have shown a great amount of promise to be used as an electronically administered resource directed by caregivers for risk evaluation at a preliminary stage or a pre-clinical screening along with triage that assists in increasing the speed of its diagnosis.

The US National Institutes of Health (NIH) has funded a repository of research data that was created by an integration of heterogeneous datasets that share agreements among researchers of autism using agreements of data sharing and this was known as the National Database for Autism Research (NDAR). Payakachat et al [17], had considered the NDAR as the largest repository for genomic data and neuroscience in autism research. Additionally, aside from biomedical data, the NDAR consists of a large collection of behavioural and clinical assessments with outcomes of health from several interventions that were novel. More importantly, the NDAR had a unique and global patient identifier which was linked to the individual data for the generation of hypothesis and testing to replicate findings of the research. The NDAR further promotes collaboration maximizing public investment and the technologies of screening and intervention for autism can be very expensive. Owing to this, the Health Services Research (HSR) with the Health Technology Assessment (HTA) aim at generating some more evidence in order to facilitate such implementation. The article further describes the NDAR and also explains its value to the researchers of health services and the decision scientists who are interested in such conditions of mental health. In Table 1, the comparisons for literature are shown.

Parikh and He [18] aimed at the exploration of the power of prediction of the Personal Characteristic Data (PCD) taken from a dataset which is well-characterized for improving the models of diagnostics of ASD. There were six personal traits (age, handedness, sex, and three other individual IQ measures) extracted from a total of 851 subjects found in the database of Autism Brain Imaging Data Exchange (ABIDE). This was an international project to collect data from many patients of ASD with typical and non-ASD controls obtained from a total of 17 clinical and research institutes. This database available publicly was employed for testing a total of 9 models of supervised machine learning. There was another strategy of cross-validation that was employed for training and further testing the models of machine learning to differentiate between the typical and Non-ASD controls along with the ASD patients. The performance of such classification was assessed by parameters like sensitivity, specificity, accuracy, and area falling under the Receiver Operating Characteristic Curve (AUC). Among the 9 different models that were tested with 6 personal traits, the model of neural network was found to be the best in terms of performance having a AUC (SD) of about 0.646 (0.005), which was followed by the K-Nearest Neighbour having a mean AUC (SD) of about 0.641 (0.004). The study was able to establish an optimal classification of ASD along with PCD as the features. By using the additional features that were discriminative such as neuroimaging, the models of machine learning could enable autism diagnosis.

Büyükoğlu and Öztürk [19] had performed some comparisons with different methods of classification like the Random Forest, on UCI 2017 Autistic Spectrum Disorder Screening Data for the Children dataset, the RBFN (Radial Basis Function Network), the IBk (K-Nearest Neighbours) and the Naïve Bayes.

There are several clinicians who are inexperienced and not confident in cases of autism since their diagnosis and calculated grade is not uniform. The availability of experts to provide expert diagnosis may also be challenging. Thus, there may be a need for a system assisted by computers to advance the power of the methods of diagnosis. This system helps in confirming the assessment of the clinicians. The works of research on the techniques of Machine Learning and the development of assessment and grading of autism is needed. Another approach using techniques of machine learning to enable grading of autism was proposed by Kanimozhiselvi et al [20].

### III. Conclusion

Machine Learning refers to a group of techniques within data distribution in order to make new decisions on data. The work is helpful in identifying aspects of different literature that were based on techniques of Big Data and Machine learning for persons with autism of Down Syndrome. All merits of this review are listed in the comparison table. Different techniques are employed to help people that are affected by Down Syndrome and Autism. The diagnosis and the treatment can be made in the early stages using machine learning and data analytics. Various techniques has helped in diagnosis, further work related to feature selection, optimal classifiers needs to be explored.

### References

- [1] Cai, H., Zhao, H., Liu, Y., & Li, G. (2018, May). Research on application of healthcare data in big data era. In *2018 International Conference on Robots & Intelligent System (ICRIS)* (pp. 377-379). IEEE.
- [2] Al-jawahiri, R., & Milne, E. (2017). Resources available for autism research in the big data era: a systematic review. *PeerJ*, 5, e2880.
- [3] Feng, B., Hoskins, W., Zhou, J., Xu, X., & Tang, J. (2017, June). Using supervised machine learning algorithms to screen down syndrome and identify the critical protein factors. In *International Conference*

- on *Intelligent and Interactive Systems and Applications* (pp. 302- 308). Springer, Cham.
- [4] Kulan, H., & Dag, T. (2018, October). Using Machine Learning Classifiers to Identify the Critical Proteins in Down Syndrome. In *Proceedings of the 2018 2nd International Conference on Computational Biology and Bioinformatics* (pp. 51-54). ACM.
- [5] Koivu, A., Korpimäki, T., Kivelä, P., Pahikkala, T., & Sairanen, M. (2018). Evaluation of machine learning algorithms for improved risk assessment for Down's syndrome. *Computers in biology and medicine*, 98, 1-7.
- [6] Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. (2019). Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. *Review Journal of Autism and Developmental Disorders*, 6(2), 128-146.
- [7] Ramanathan, S., Sangeetha, M., Talwai, S., & Natarajan, S. (2018, September). Probabilistic Determination Of Down's Syndrome Using Machine Learning Techniques. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 126-132). IEEE.
- [8] Çelik, E., İlhan, H. O., & Elbir, A. (2017, May). Detection and estimation of down syndrome genes by machine learning techniques. In *2017 25th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
- [9] Furat, F. G., & Ibrikci, T. Classification of Down Syndrome of Mice Protein Dataset on MongoDB Database. *Balkan Journal of Electrical and Computer Engineering*, 6(2), 44-49.
- [10] Feng., L. "Big Data Application in Genetic Studies of Down syndrome".
- [11] Lombardo, M. V., Lai, M. C., & Baron-Cohen, S. (2019). Big data approaches to decomposing heterogeneity across the autism spectrum. *Molecular psychiatry*, 1.
- [12] Obafemi-Ajayi, T., Lam, D., Takahashi, T. N., Kanne, S., & Wunsch, D. (2015, August). Sorting the phenotypic heterogeneity of autism spectrum disorders: A hierarchical clustering model. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-7). IEEE.
- [13] Rudovic, O., Lee, J., Dai, M., Schuller, B., & Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19), eaao6760.
- [14] Radwan, A. M., Birkan, B., Hania, F., & Cataltepe, Z. (2017). Active machine learning framework for teaching object recognition skills to children with autism. *International Journal of Developmental Disabilities*, 63(3), 158-169.
- [15] Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders*, 45(5), 1121-1136.
- [16] Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Translational psychiatry*, 6(2), e732.
- [17] Payakachat, N., Tilford, J. M., & Ungar, W. J. (2016). National Database for Autism Research (NDAR): Big data opportunities for health services research and health technology assessment. *PharmacoEconomics*, 34(2), 127-138.
- [18] Parikh, M. N., Li, H., & He, L. (2019). Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data. *Frontiers in computational neuroscience*, 13.
- [19] Büyükoğluz, F. N., & Öztürk, A. (2018, May). Early autism diagnosis of children with machine learning algorithms. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
- [20] Kanimozhiselvi, C. S., Jayaprakash, M. D., & Kalaivani, M. K. (2019). Grading Autism Children Using Machine Learning Techniques. *International Journal of Applied Engineering Research*, 14(5), 1186-1188.