Data Imbalance Impact on Autism Pre-diagnosis System: An Experimental Study

Neda Abdelhamid
IT Programme
Auckland Institute of Studies
Auckland, New Zealand
nedah@ais.ac.nz

Arun Padmavathy
Digital Technologies,
Manukau Institute of
Technology
Auckland, 2241, New
Zealand
jana12@manukaumail.com

David Peebles
School of Health,
Psychology Dept.
University of Huddersfield
Queensgate, Huddersfield
HD1 3DH, UK
d.peebles@hud.ac.uk

Fadi Thabtah
Psychology Department
University of Huddersfield
Huddersfield, UK
f.thabtah2@hud.ac.uk

Daymond Goulder-Horobin
Digital Technologies, Manukau Institute of Technology
Auckland, 2241, New Zealand
goul86@manukaumail.com

Abstract

Machine Learning (ML) is a branch of computer science that is rapidly gaining popularity within the healthcare arena due to its ability to explore large datasets to discover useful patterns that can be interepreted for decision making and prediction. ML techniques are used for the analysis of clinical parameters and their combinations for prognosis, therapy planning and support, and patient management and wellbeing. In this research, we investigate a crucial problem associated with medical applications such as Autism Spectrum Disorder (ASD) data imbalances in which cases are far more than just controls in the dataset. In autism diagnosis data, the number of possible instances is linked with one class, i.e. the No ASD is larger than the ASD, and this may cause performance issues such as models favouring the majority class and undermining the minority class. This research experimentally measures the impact of class imbalance issue on the performance of different classifiers on real autism datasets when various data imbalance approaches are utilized in the pre-processing phase. We employ Oversampling techniques such as Synthetic Minority Oversampling (SMOTE), and Undersampling with different classifiers including Naive Bayes, RIPPER, C4.5, and Random Forest to measure the impact of these on the performance of the models derived in terms of Area Under Curve (AUC) and other metrics. Results pinpoint that Oversampling techniques are superior to Undersampling techniques, at least for the toddlers' autism dataset that we consider, and suggest that further work should look at incorporating sampling techniques with feature selection to generate models that do not overfit the dataset.

Keywords: *Autism Spectrum Disorder, ASD Screening, Data Imbalance, Machine Learning, Undersampling, Oversampling, SMOTE*

1 Introduction

ASD is a neurodevelopmental disorder referring to impairment in social interaction, verbal and nonverbal language, interests, activities, and a stereotyped restricted way of behaviour (Thabtah, Abdelhamid & Peebles, 2019; Parellada, et al., 2014). These symptoms are generally seen in the early childhood period, but also in older children and adolescents in some parts of the world. Children with autism do not share their enjoyment of objects or pointing at and showing things to others (Willemsen-Swinkels & Buitelaar, 2002). Often their facial expressions, do not reflect the things they are saying. They often avoid making eye contact, and have trouble with voice tone and not understanding the emotions and intentions of others (Madipakkam, Rothkirch, Dziobek, & Sterzer, 2017). Restrictive and repetitive behaviours are also common such as rocking their body, lining up or spinning objects, and staring at lights, etc (Ravizza, Solomon, Ivry, & Carterb, 2017).

Preliminary treatment of ASD can reduce further development and provide rapid access to necessary healthcare resources (Thabtah, 2018a). ASD is typically diagnosed by psychiatrists and clinicians using certain criteria as desseminated in the Diagnostic and Statistical Manual of Mental Disorders Revision 5 (DSM-V) (American Psychiatric Association, 2013). Individual strengths and weaknesses in the area such as activities, behaviour, repetitive interest, social communication, sensory processing, and social interaction are assessed during the diganosis process in a clinical setting (Thabtah, 2017). This diagnosis involves two steps according to Allison, et al., 2008; the first step is screening, in which parents answer a series of questions relating entirely to behavioural traits of the child (mental, physical, etc). The second stage is a comprehensive diagnostic evaluation conducted by a multidisciplinary team which gathers data using different methods like interviews and observations based on DSM-V criteria.

ML provides techniques which are useful in prognostic and diagnostic problems in a variety of medical fields. The clinical parameters can be analyzed with ML and their combination for prognosis; for example, extraction for medical knowledge for outcome research, prediction of disease progression, therapy planning and support, and overall patient management. ML is also a part of data analysis; it helps to find irregularities in the data and then interpretation of the data is used in the Intensive Care Unit and Intelligent Alarming offering rich results in an effective and efficient manner (Vellido, Ribas, Morales, Sanmartín, & Rodríguez, 2018). In the healthcare environment, successful implementation of ML methods can help the integration of computer-based systems, improving the work of medical experts, and providing an opportunity to facilitate and improve the quality and efficiency of medical care (Achenie et al., 2019; Thabtah & Peebles, 2019).

Health applications, such as autism diagnosis and screening, are often associated with uneven datasets, in which one of the classes has low frequency in the dataset (Chawla, Bowyer, Hall, & Kegelmeye, 2002; Fernández, Río, Chawla, & Herrera, 2017; Abdeljaber, 2019). In most ML algorithms, the classifiers implicitly assume that the training dataset is balanced, therefore standard classifiers usually derive models that are biased towards the majority class (Estabrooks, Jo, & Japkowicz, 2004). ASD screening is a classic classification problem that employs a supervised learning process to predict whether individuals exhibit autistic traits, hence a labelled training data with somewhat balanced class labels is needed. In autism screening, if the training dataset is imbalanced this may cause high misclassification especially in terms of false negatives (individuals who are predicted not to be on the spectrum when they actually are). These misclassifications may delay access to medical resources and cause a long-term delay for individuals who are urgently needing medical intervention (speech therapy, special education, medical care, etc).

During the training phase of the classifier on an uneven dataset, the focus is usually on the majority (instances that are not on the spectrum); the classifier may overlook the minority class (instances which are on the spectrum) and conventional evaluation metrics, such as error rate and accuracy, produce biased results. Therefore, it is imperative to deal with the class imbalance issue either in the preliminary phases (data-related treatments) or during building the classification model (algorithms treatments) to maintain adequate models with a non-biased performance. Data balancing can be employed using oversampling or undersampling techniques (Chawla, Bowyer, Hall, & Kegelmeye, 2002).

This research investigates the impact of class imbalance of ASD screening on the performance of ML models derived by classifiers. In particular, we utilise Oversampling and Undersampling techniques and contrast their performance on an autism classification system derived by different classification algorithms such as Naïve Bayes, Random forest, RIPPER and Decision Tree (C4.5) (Duda & Hart, 1973; Breiman, 2001; Cohen, 1995; Quinlan, 1993). The research question we are trying to answer is:

Can data imbalance sampling methods improve the classification performance of autism detection in toddlers with respect to AUC and other evaluation metrics?

We try to ascertain which sampling techniques provide the best predictive performance for autism detection, if any, by contrasting a number of data imbalance sampling techniques including SMOTE, Random Oversampling (ROS) and Random Under-Sampling (RUS) among others on real datasets related to toddlers (Chawla et al., 2002; Zhuoyuan, Yunpeng, & Ye, 2015; Kubat & Matwin, 1997).

The structure of this research includes the problem statement in Section 1 which is followed by literature reviews on data imbalance and ASD screening using ML in Section 2. In Section 3 a data description of ASD dataset follows which includes details of the attributes of the dataset. Experimental result analysis ensues in the subsequent discussion in Section 4, and finally, conclusions are given in Section 5.

2. The Problem and Literature Review

2.1 The Problem

Data imbalances cause a significant problem in classification as the performance of classifiers processing uneven datasets is often biased towards the majority class (Yang & Wu, 2006). Datasets related to autism screening are no exception as they are normally imbalanced with respect to the response variable because instances are normally linked with many more controls than cases (Thabtah, 2017; Thabtah 2018a). So, during the building of the model, the classifiers often undermine the minority class (having ASD), which decreases the performance of the classifier when processing cases and controls to build classification models. Figure 1 depicts autism screening and pinpoints the primary phases needed to deal with the class imbalance issue.

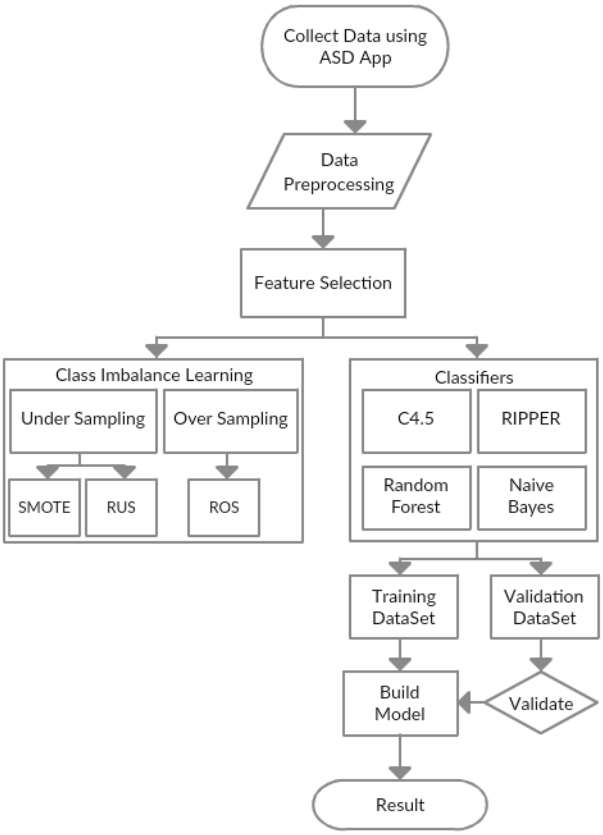

Figure 1. ASD Screening as a class imbalance Problem

2.2 Background on Sampling Techniques

Random Oversampling is a method in which the minority class in the sample is duplicated to the point where the class attribute is balanced. This is the baseline method of oversampling a dataset, and the literature suggests that while this is the simplest method, it often creates heavy overfitting in the dataset (Zhuoyuan, Yunpeng, & Ye, 2015).

RUS resamples the dataset while ensuring that the class attribute is balanced. The issue with this method is there is a potential loss of information if a significant amount of undersampling needs to be done, especially in medical diagnosis to some extent (Shelke, Deshmukh, & Shandilya, 2017).

SMOTE) is one of the most common methods of oversampling a dataset aside from the random method (Chawla et al., 2002). It works by using the K Nearest Neighbours (KNN) clustering technique to cluster the dataset; this will randomly generate synthetic examples between the neighbours in the sample rather than replicating the instances to generate more realistic samples. For this project, because the dataset has largely categorical variables, we have chosen to use SMOTE-NC which is designed for both numeric and categorical attributes with a nominal class variable.

2.3 Literature Review

Rahman and Davis (2013), and Li, Liu, and Hu, (2010) compared ROS, and the SMOTE techniques to determine which offers better performance. Rahman and Davis (2013) applied the two techniques in a trial-and-error fashion against a Modified Cluster-Based Under-Sampling Method. Liu, and Hu, (2010) investigated undersampling to be able to compare, and then to present the superior classification methods. The authors have used different datasets related to supervised learning applications.

Dittman, Khoshgoftaar, Wald, and Napolitano (2014) developed a ML method to classify genes based on cancerous and non-cancerous, with datasets having a class imbalance with the minority class at 35%. The dataset used was from real world bioinformatics, genetics and medical areas with more than 25 features. The authors used two classifiers: K Nearest Neighbour, and Support Vector Machines (Hall, Park, & Samworth, 2008; Cortes & Vapnik, 1995) to build the model. After applying sampling techniques, the results showed that RUS outperforms SMOTE and ROS. But the difference between the best and the worst sampling techniques, comparing AUC is ≤ 0.01, and so there is little statistical difference between the sampling techniques. Henceforth, the authors recommended RUS as the preferred sampling technique due to less computational cost than SMOTE.

Dubey, Zhou, Wang, Thompson, and Ye (2013) investigated the class imbalance problem in Alzheimer’s Disease by using a Neuroimaging Initiative dataset. The dataset was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (Initiative, n.d.), which consists of MRI images and features. The two classifiers used included Random Forest and Support Vector Machines (Breiman, 2001; Cortes & Vapnik, 1995) and the performance metrics used were AUC, sensitivity, and specificity measures. The results showed that the best performance was attained with the undersampling technique based on K-Medoids technique, which is a variant of the k-means algorithm.

El-Sayed, Meguid, Mahmood, and Hefny (2015) applied SMOTE imbalanced data on autism collected from the National Research Centre in Egypt and attained higher performance. The dataset consisted of data acquired from 36 questionnaires on social, communication, and restricted behaviours of adults. The data was imbalanced with 100 people with autism belonging to the majority class, and 15 non-autistic in the minority class. The classifiers used were Sequential Minimal Optimization (SMO), C4.5 and Naïve Bayes and Multi-Layer Perception Neural Network (Platt, 1998; Quinlan, 1993; Duda & Hart, 1973; Haykin, 1999). The results showed that oversampling of imbalanced data makes the model more reliable. Since the data used in the study is limited, more cases and controls are needed and further testing is required.

In this research, a SMOTE was presented for handling autism-imbalanced data to increase accuracy credibility. SMOTE can potentially lead to over-fitting on multiple copies of minority class examples.

3. Data and Pre-Processing Phase

Table 1: Description of the Toddler Dataset Attributes

Attribute	Description	Type
A1	Does your child look at you when you call his/her name?	Binary
A2	How easy is it for you to get eye contact with your child	Binary
A3	Does your child point to indicate that he/she wants something?	Binary
A4	Does your child point to share interest with you?	Binary
A5	Does your child pretend?	Binary
A6	Does your child follow where you're looking?	Binary
A7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them?	Binary
A8	How would you describe your child's first words?	Binary
A9	Does your child use simple gestures?	Binary
A10	Does your child stare at nothing with no apparent purpose?	Binary
Age	Age of toddler	Numeric
Score by Q-CHAT-10	1-10 (Less than or equal 3 no ASD traits, greater than 3 ASD traits)	Numeric
Sex	Male or Female	Character
Ethnicity	List of common ethnicities in text format	String
Born with Jaundice	Whether the case was born with jaundice	Boolean
Family Member with ASD history	Whether any immediate family member has a PDD	Boolean
Class	Whether individuals exhibit ASD traits	Boolean

The dataset we are using for our analysis is retrieved from ASDTest, (2017) and was donated by Thabtah, (2018b). The dataset consists of 18 attributes and 1054 instances. The attributes A1–A10 are ten questions based on a conventional medical screening questionnaire called Quantitative Check List for Autism in Toddler (Q Chat-10) (Allison, Auyeung, & Baron-Cohen, 2012) (See Table 1). The answers to these questions are assigned as 0 and 1 based on respondents' answers. (Details on data transformation can be found in Thabtah & Peebles and Thabtah, Kamalov & Rajab, (2018)). Score attribute is the total score obtained after adding all points for the questions from A1–A10; if the score is more than three, the class value is assigned 'Yes' depicting the toddler has ASD, and if the score is less than three the value of the class variable is 'No' indicating that the toddler has no ASD traits. The attribute Family ASD is historical data which tells whether any of the family members of the child has any autism history.

The dataset is imbalanced with the majority class 'Yes' (toddlers having ASD) of 728 occurrences, and minority class being 'No' (toddlers having No ASD) with 326 occurrences. The classifier may be biased towards classifying a screening with majority class 'Yes'. 'No' is the class that will be over-sampled during the pre-processing step. To ensure we had fair training by the classifiers, we removed the score attribute from the dataset, as the class attribute is directly derived from the score.

3.1 Data Pre-processing

Firstly, the SMOTE technique in WEKA (Waikato Environment for Knowledge Analysis) was applied using the SMOTE package to over-sample the minority class so it would equal the majority class, and make the class attribute balanced. For this, the traditional setting of K-5 was assigned, which helps to find the number of K clusters to generate the synthetic samples in the sample.

We did have some issues using SMOTE. It generated samples with a score three or greater which would have been categorized as 'Yes' in the original dataset, potentially causing issues. Secondly, as the data has nominal values, the SMOTE may not have worked as efficiently as possible. Even though the package in WEKA can handle this, it may not be as good. We also created subsets using the WEKA tool Spread Sub Sample and described the settings to ensure that the class balance was 1:1. This decreased the sample size but made the class balanced and gave us an undersampled dataset for the analysis. We then created the randomly oversampled dataset using the knowledge flow module in WEKA and validated the dataset using visualization.

4. Experimental Analysis

A. Experimental Settings

For this research, we have used WEKA 3.8 which was developed by the University of Waikato in New Zealand (Hall, et al., 2009). It is an open-source, Java-based tool that consists of a collection of data pre-processing and machine algorithms organised inside packages. This tool is used for a variety of tasks such as regression, clustering, association, data pre-processing, classification and limited visualization. We used the WEKA explorer module to run the classifiers, and the Knowledge Flow module to run the sampling techniques.

Experiments were run on a Windows 10 Operating system with an Intel i7-6700HQ, 2.6GHZ of power and 16GB of RAM.

B. Classification Algorithms Used

We have used classification techniques in the experiments that adopt various learning schemes, hence ensuring fair results analysis. The first classification technique that we used was the C4.5 (Quinlan, 1993) which is a type of Decision Tree algorithm. A conventional Decision Tree algorithm performs based on the information gain metric and displays results through a Hierarchical Tree Model of Decisions and Outcomes. A Decision Tree can be good for predicting a target variable and new data. Random Forest (Breiman, 2001) is another classification technique we used that randomly generates Decisions Trees from random features and uses the most accurate ones to produce its results. It can be used for classification and regression problems.

We have also used Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Cohen, 1995), which employs a propositional rule learning scheme to produces a set of logic statements with a set of attributes and an outcome of the class. Lastly, Naïve Bayes (Duda & Hart, 1973) has been adopted which is a type of probabilistic classifier that uses the independence of the attribute's assumption to predict test data based on observed probabilities of the test data attributes' values within the training dataset.

C. Evaluation Metrics Used

To evaluate the performance of our model, we have used the Confusion Matrix (Fawcett, 2006) (Table 2). Sensitivity (SN) (Equation 1) is determined as the quantity of correct positive predictions divided by the total number of positives. Specificity (SP), also known as the True Negative Rate (TNR) (Equation 2), correctly identifies the number of negatives from the total number of negatives. Likewise, Precision (PR) (Equation 3) is determined as the quantity of correct positive predictions divided by the absolute number of positive predictions. It is additionally called positive predictive value (PPV).

Area Under Curve – Precision and Recall (AUC-PR) (Equation 4) is one of the main measures we will be using, rather than accuracy, for evaluating the models due to the class imbalance problem. The AUC-PR uses both the precision and recall statistics and plots it on a curve. As the SN increases, the PR can only decrease as more weighting is put on the positive class based on various thresholds. False Positive Rate (FPR) (Equation 5) denotes the quantity of inaccurate positive predictions partitioned by the complete number of negatives.

$$\text{Sensitivity (SN)} = \frac{\text{TP}}{\text{TP+FN}} \qquad (1)$$

$$\text{Specificity (SP) (TNR)} = \frac{\text{TN}}{\text{TN+FN}} \qquad (2)$$

$$\text{Precision (PR)/PPR} = \frac{\text{TP}}{\text{TP+FP}} \qquad (3)$$

$$\text{AUC-PR} = \frac{P-R+1}{2} \qquad (4)$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP+TN}} \qquad (5)$$

Matthews Correlation Coefficient (MCC) (Equation 6) is used in ML evaluation to measure the quality of a binary classifier. The value can range from -1, which indicates a complete disagreement between the predicted and the actual values, and 1 which indicates perfect prediction. MCC is generally considered to be more informative than accuracy, as it takes into account the balance of the confusion matrix between the FP, FN, TP and TN and thus is more useful with an imbalanced class attribute.

$$\text{MCC} = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (6)$$

4.2 Results Analysis

Tables 3.1–3.4 depict results derived by the different classification algorithms from the original toddler dataset and from different data versions after applying SMOTE, ROS, and RUS techniques. Based on the figures in the tables, it can be seen that the AUC-PR rate has increased significantly for all the classification models when sampling techniques have been applied to the autism dataset. In particular, models derived by the Naïve Bayes classifier showed superiority in terms of AUC-PR with 99.80% in predicting autistic traits, especially when the SMOTE technique was applied in the pre-processing phase. Similarly, ROC results have also increased after applying sampling techniques. The specificity rate derived by the Naïve Bayes classifier was 1.0 when Random with RUS and ROS were applied in the pre-processing phase, which pinpoints the impact uneven class labels may have on autism screening systems.

Table 2: Confusion Matrix for ASD Screening Problem

	Predicted Class Value	
Actual Class Value	**ASD**	**No-ASD**
ASD	True Positive (TP)	False Negative (FN)
No-ASD	False Positive (FP)	True Negative (TN)

Table 3.1 Metric Comparison of Original Dataset

Classifier	Original				
	SN	SP	F-Measure	AUC-ROC	AUC-PR
C4.5	92.30%	93.70%	92.30%	92.50%	90.80%
Random Forest	95.30%	96.00%	95.20%	99.20%	99.30%
RIPPER	92.40%	95.80%	92.50%	94.80%	94.50%
Naïve Bayes	96.20%	99.40%	96.30%	99.70%	99.70%

Table 3.2 Metric Comparison After Applying SMOTE

Classifier	SMOTE				
	SN	SP	F-Measure	AUC-ROC	AUC-PR
C4.5	94.30%	93.80%	94.30%	94.70%	92.20%
Random Forest	97.00%	97.30%	97.00%	99.70%	99.70%
RIPPER	94.80%	93.60%	94.80%	97.20%	96.30%
Naïve Bayes	98.80%	97.00%	97.00%	99.80%	99.80%

The harmonic mean of Precision and Recall (F-Measure) (Equation 7) achieved its highest value of .970 when SMOTE was applied with the Naïve Bayes classifier on the toddler dataset. While calculating F-measure, the confusion matrix evaluation metrics TP, TN, FP, and FN are considered. In ASD screening, non-recognition of ASD or FN can lead to a disastrous error rather than FP error which can be corrected by re-diagnosis. Since the F-measure considers both Precision and Recall, it is a relatively better performance metric. For all the classifiers applied after sampling methods, the F-measure results derived by their models have shown an increase of 2% on an average when compared to F-Measure rates obtained by the same classifiers from the original dataset and without sampling.

$$\text{F-Score} == 2 \text{ X } \frac{\text{Precision X Recall}}{\text{Precision+ Recall}} 2 \qquad (7)$$

The AUC-ROC curve obtained by plotting TPR against FPR has also shown an increase in 2.3% for the RIPPER classifier while applying SMOTE on the toddler dataset. Interestingly, all the values for AUC-ROC by the classifiers were found to be >95% when using SMOTE and ROS techniques; this highlights that the models are statistically sound. If we rank the results of the AUC-ROC for the same set of learners and samplers, it can be seen that RIPPER always derived higher AUC-ROC except in one case when RUS was applied in the pre-processing phase. This reveals that rule induction approaches such as RIPPER sare less superior than a simple probabilistic classifier such as Naïve Bayes at least when sampling techniques were applied at the toddler autism dataset. This also indicates that the RIPPER may be sensitive to class imbalances especially in distinguishing between the positive and negative class labels. Similarly, the results of AUC-PR & AUC-ROC for all learners and samplers were similar except when RUS was applied. One interesting note based on the results obtained is that the sampled dataset's results were different from those obtained by the same classifiers on the original dataset, which reveals how vital it is to balance the data before learning models, at least on the autism screening application we consider.

We tested two additional classifiers (k-nearest neighbors – KNN & PART) (Cover and Hart, 1967; Frank & Witten, 1998) to verify earliest classifications results when using resampling mehtods and in terms of SP and SN rates. The results obtained againist the different resampling datasets revealed that SP rate of both KNN and PART are at maximum when ROS method was used to pre-process the autism dataset. Moreover, the SN rate produced by PART classifier is also the highest when ROS was used to pre-process the dataset. However, the SN rate when KNN classifier is derived was the highest when SMOTE resampling method was used. Nevertheless, results clearly pinpointed consistency that oversampling methods have indeed imrpoved the perfornance when sampled the autism dataset regardless the type of classifiers used at least on the classification algorithms we cosnidered in this research.

Table 3.3 Metric Comparison after applying RUS

Classifier	RUS				
	SN	SP	F-Measure	AUC-AUC	AUC-PR
C4.5	92.30%	93.30%	92.30%	93.00%	90.10%
Random Forest	94.50%	92.90%	94.50%	99.30%	99.30%
RIPPER	92.20%	93.30%	92.20%	94.10%	92.00%
Naïve Bayes	96.90%	99.90%	96.90%	99.60%	99.70%

Table 3.4 Metric Comparison After Applying ROS

Classifier	ROS				
	SN	SP	F-Measure	ROC-AUC	PR-AUC
C4.5	94.80%	96.80%	94.80%	96.70%	95.70%
Random Forest	97.50%	98.50%	97.50%	99.90%	99.90%
RIPPER	95.50%	96.80%	95.50%	97.30%	96.40%
Naïve Bayes	96.40%	99.90%	96.40%	99.70%	99.70%

Table 3.5 MCC Results Obtained by the Learners

Mathews Correlation Coefficient	Original	SMOTE	ROS	RUS
C4.5	82.70%	87.60%	91.00%	81.60%
Random Forest	88.80%	93.70%	95.20%	89.00%
RIPPER	86.70%	91.60%	93.60%	85.00%
Naïve Bayes	91.80%	94.00%	93.10%	94.00%

We have included the MCC in the results (Table 3.5) to show the effectiveness of the learnt autism screening models. The derived results of the MCC show that the best models in terms of MCC have been obtained when the combination of ROS and Random Forest classifier have been used. However, when the original dataset was processed without any sampling, Naïve Bayes maintained superiority over the remaining classifiers. Overall, when sampling techniques were applied, most classifiers produced good MCC results when compared to processing the original dataset without sampling except in one case, i.e. when RIPPER was applied with RUS.

Overall, KNN, PART, Naïve Bayes and Random Forest classifiers appear to work well with the SMOTE and ROS sampling methods to deal with the imbalanced issue on the toddler dataset; this was consistent for most evaluation metrics results. Most results obtained, according to the evaluation metrics, show that sampling techniques have the performance of the classifiers when compared to those obtained without sampling, i.e. original dataset. Based on the results generated, Random Forest with ROS produced the best model with reasonable balance, though not a perfectly balanced classifier. If wanting to compromise on accuracy with a more balanced classifier, then we recommend Random Forest with SMOTE which will produce a balanced classifier. The Naive Bayes classifier has increased the reliability of classification while applying the sampling techniques. Using SMOTE, there were only 43 misclassified instances, while using ROS misclassified 53 instances despite that the FPR of oversampled data was 0. It is notable that the FPR and FNR have significantly reduced when sampling techniques are applied on the toddler autism dataset. Finally, the AUC-PR for the best two classifiers when SMOTE was used have been depicted in Figure 2.

Table 4 KNN and PART algorithms performance using different resampling methods

KNN Classifier		
Resampling Method	**SN**	**SP**
RUS	95.10%	92.41%
ROS	96.60%	99.74%
SMOTE	98.62%	94.26%
No Resampling	96.76%	88.59%
PART Classifier		
Resampling Method	**SN**	**SP**
RUS	96.83%	94.04%
ROS	96.70%	98.29%
SMOTE	97.47%	96.01%
No Resampling	96.96%	92.54%

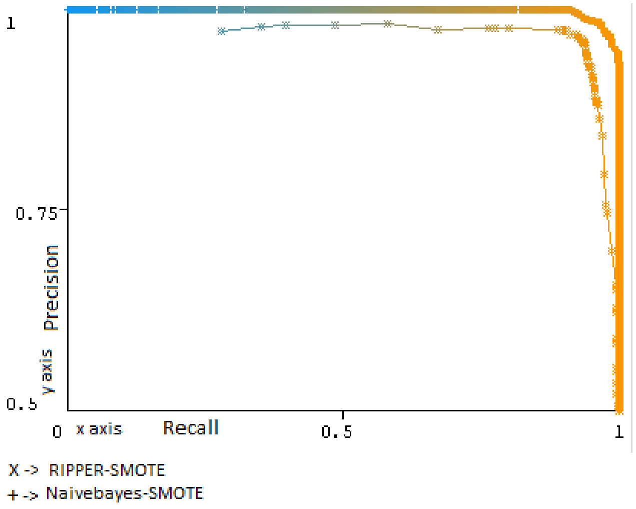

Figure 2: Comparison of AUC-PR for RIPPER and Naive Bayes Classifiers with SMOTE

5 Conclusions

Class imbalance is one of the most persistent problems when dealing with medical and bioinformatics datasets since the classifiers tend to increase accuracy which is biased on the majority class. Data sampling techniques are used to overcome this problem in which instances are added or removed from the existing dataset to adjust the class ratio. In this research, we investigate one crucial behavioural application that often contains imbalance class - autism, to reveal the true classification performance of classifiers in the presence of class imbalance. To achieve the aim, different sampling methods (SMOTE, ROS, RUS) with four classifiers (Naïve Bayes, C4.5, RIPPER, Random Forest) have been experimentally contrasted to pinpoint the impact of sampling on autism screening applications. The results derived from a real screening dataset relating to toddlers, with respect to different evaluation metrics, showed ROS helps classifiers to derive superior results with respect to SN, SP, AUC, and F-score followed by SMOTE, at least on the toddler dataset. The overall results pinpoint that sampling techniques improve the classifier's performance and ensured balanced classifiers have been obtained through the evaluation of the models using AUC-PR and AUC-ROC. However, the best sampling method to apply will depend on which classifier has been employed as there was some result variations. For instance, if Random Forest is utilised, then ROS is the best sampling to use; if Naïve Bayes is used, then SMOTE seems more appropriate for the dataset we considered.

There are many ways to move forward, such as it may be good to compare the variations of SMOTE, to see what impact it has had on many different versions of the autism dataset. There is also an opportunity to explore further cost-based classification, which was not covered in this research, or to consider Boosting methods like ADABoost and LogitBoost, to have them weigh the class attributes and see how they handle the original as well as potentially sampled datasets. .

References

[1] Abdeljaber Thabtah F. (2019) Detecting Autistic Traits using Computational Intelligence & Machine Learning Techniques. Master of Research thesis, Psychology Department, School of Health, University of Huddersfield, 2019.

[2] Achenie, L., Scarpa, A., Factor, R., Wang, T., Robins, D., & McCrickard, D. (2019). A Machine Learning Strategy for Autism Screening in Toddlers. *J Dev Behav Pediatr*. doi:10.1097/DBP.0000000000000668

[3] Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in toddlers in 1,000 cases and 3,000 controls. *J*

Am Acad Child Adolesc Psychiatry. doi:10.1016/j.jaac.2011.11.003

[4] Allison, C., Baron-Cohen, S., Wheelwright, S., Charman, T., Richler, J., Pasco, G., & Brayne, C. (2008). The Q-CHAT (Quantitative CHecklist for Autism in Toddlers):A Normally Distributed Quantitative Measure of Autistic Traits at 18–24 Months of Age: Preliminary Report. *J Autism Dev Disord*. doi:10.1007/s10803-007-0509-7

[5] Association, A. P. (2013). *Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition.* United States. doi:https://doi.org/10.1176/appi.books.9780890425596

[6] Breiman, L. (2001). Random Forests. *Machine Learning*. doi:https://doi.org/10.1023/A:1010933404324

[7] Chawla, N., Bowyer, K., Hall, L., & Kegelmeye, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *JAIR*. doi:https://doi.org/10.1613/jair.953

[8] Cohen, W. W. (1995). Fast Effective Rule Induction. *Machine Learning Proceedings of the Twelfth International Conference.*

[9] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*. doi:10.1007/BF00994018

[10] Cover T and Hart P (1967). Nearest neighbor pattern classification. *IEEE Trans Inform Theory* 13(1): 21–27

[11] Dittman, D. J., Khoshgoftaar, T. M., Wald, R., & Napolitano, A. (2014). Comparison of Data Sampling Approaches for Imbalanced Bioinformatics Data. *The Twenty-Seventh International Flairs Conference.* AAAI Publications.

[12] Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2013). ANALYSIS OF SAMPLING TECHNIQUES FOR IMBALANCED DATA: AN N=648 ADNI STUDY. doi:10.1016/j.neuroimage.2013.10.005

[13] Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley-Interscience publication.

[14] El-Sayed, A. A., Meguid, N. A., Mahmood, M. A., & Hefny, H. A. (2015). Handling Autism Imbalanced Data using SyntheticMinority Over-Sampling Technique (SMOTE). *2015 Third World Conference on Complex Systems (WCCS).* doi:10.1109/ICoCS.2015.7483267

[15] Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*. doi: https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

[16] Fernández, A., Río, S. d., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced Big Data classification: outcomes and challenges. *Complex & Intelligent Systems*. doi:https://doi.org/10.1007/s40747-017-0037-9

[17] Frank F., Witten I. H. (1998) Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* doi:10.1145/1656274.1656278

[19] Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* Prentice Hall.

[20] Initiative, A. D. (n.d.). *http://adni.loni.usc.edu/data-samples/adni-data-inventory/*. Retrieved from http://adni.loni.usc.edu/.

[21] Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *International Conference on Machine Learning* (pp. 179-186). Nashville, Tennesse: Morgan Kaufmann.

[22] Li, D., Liu, C., & Hu, S. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine, 40*(5), 509-518.

[23] Madipakkam, A. R., Rothkirch, M., Dziobek, I., & Sterzer, P. (2017). Unconscious avoidance of eye contact in autism spectrum disorder. *Scientific Reports*. doi:10.1038/s41598-017-13945-5

[24] Parellada, M., MJ, P., L, P., C, M., E, G.-V., G, Z., & C, A. (2014). The neurobiology of autism spectrum disorders. *Eur Psychiatry*. doi:10.1016/j.eurpsy.2013.02.005. Epub 2013 Nov 22

[25] Platt, J. C. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.* Microsoft Research. doi:10.1.1.43.4376

[26] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

[27] Rahman, M., & Davis, D. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing, 3*(2), 224-228.

[28] Ravizza, S. M., Solomon, M., Ivry, R. B., & Carterb, C. S. (2017). Restricted and repetitive behaviors in autism spectrum disorders: The relationship of attention and motor deficits. *Dev Psychopathol*. doi:10.1017/S0954579413000163

[29] Shelke, M. S., Deshmukh, P. R., & Shandilya, V. K. (2017). A Review on Imbalanced Data Handling Using Undersampling and. *International Journal of Recent Trends in Engineering & Research*.

[30] Thabtah, F. (2017, November). *ASDTest*. Retrieved from Google Play: https://play.google.com/store/apps/details?id=com.asd.asdquiz&hl=en

[31] Thabtah, F. (2017). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *INFORMATICS FOR HEALTH & SOCIAL CARE*. doi:https://doi.org/10.1080/17538157.2017.1399132.

[32] Thabtah F. (2108a) Machine learning in autistic spectrum disorder behavioral research: A review and ways forward Informatics for Health and Social Care 43 (2), 1-20.

[33] Thabtah, F. (2018b). An accessible and efficient autism screening method for behavioural data and predictive analyses. *Health Informatics Journal*. doi:10.1177/1460458218796636

[34] Thabtah F, Kamalov F., Rajab K (2018) A new computational intelligence approach to detect autistic features for autism screening. International Journal of Medical Infromatics, Volume 117, pp. 112-124.

[35] Thabtah F., Peebles D. (2019) A new machine learning model based on induction of rules for autism detection. Health Informatics Journal, 1460458218824711.

[36] Thabtah F., Abdelhamid N., Peebles D. (2019) A machine learning autism classification based on logistic regression analysis. Health information science and systems 7 (1), 12.

[37] TomFawcett. (2006). An introduction to ROC analysis. *ScienceDirect*. doi:https://doi.org/10.1016/j.patrec.2005.10.010

[38] Vellido, A., Ribas, V., Morales, C., Sanmartín, A. R., & Rodríguez, J. C. (2018). Machine learning in critical care: state-of-the-art and a sepsis case study. *Biomed Eng Online*. doi:10.1186/s12938-018-0569-2

[39] Willemsen-Swinkels, S. H., & Buitelaar, J. K. (2002). The autistic spectrum: subgroups, boundaries, and treatment. *PlumX Metrics*. doi:https://doi.org/10.1016/S0193-953X(02)00020-5

[40] Yang, Q., & Wu, X. (2006). 10 Challenge Problems in Data Mining Research. *International Journal of Information Technology and Decision Making, 5*(4), 597-604.

[41] Zhuoyuan, Z., Yunpeng, C., & Ye, L. (2015). Oversampling method for imbalanced classification. *Computing and Informatics*.