İSTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

USING MACHINE LEARNING TECHNIQUES TO ENHANCE TEACHING AND PERFORMANCE PREDICTION OF STUDENTS WITH AUTISM SPECTRUM DISORDERS

Ph.D. THESIS

Akram RADWAN

Department of Computer Engineering

Computer Engineering Program

JULY 2016

İSTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

USING MACHINE LEARNING TECHNIQUES TO ENHANCE TEACHING AND PERFORMANCE PREDICTION OF STUDENTS WITH AUTISM SPECTRUM DISORDERS

Ph.D. THESIS

Akram RADWAN
(504112514)

Computer Engineering Department

Computer Engineering Program

Thesis Advisor: Prof. Dr. Zehra ÇATALTEPE

JULY 2016

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY ÖĞRENME YÖNTEMLERİ İLE OTİZM SPEKTRUM BOZUKLUĞU OLAN ÖĞRENCİLERİN ÖĞRETİMİNİN VE ÖĞRETİM PERFORMANSI TAHMİNİNİN İYİLEŞTİRİLMESİ

DOKTORA TEZİ

Akram RADWAN
(504112514)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Zehra ÇATALTEPE

TEMMUZ 2016

Akram RADWAN, a Ph.D. student of ITU Graduate School of Science Engineering and Technology, student ID 504112514, successfully defended the thesis/dissertation entitled "Using Machine Learning Techniques to Enhance Teaching and Performance Prediction of Students with Autism Spectrum Disorders", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Zehra ÇATALTEPE** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Dr. Hatice KÖSE** Istanbul Technical University	.....
	Prof. Dr. Barış KORKMAZ Istanbul University	.....
	Asst. Prof. Dr. Yusuf YASLAN Istanbul Technical University	.....
	Asst. Prof. Dr. Albert Ali SALAH Boğaziçi University	.....

Date of Submission : 20 June 2016
Date of Defense : 19 July 2016

To my parents, spouse and children,

FOREWORD

Fırst and Foremost, I would like to express my deepest gratitude to my advisor, Professor Zehra ÇATALTEPE. I benefit so much from her deep insight into machine learning, his professionalism, and her constant encouragement. Her willingness to adapt to and encourage my ever-evolving interests enabled the body of this thesis to come to fruition. Thanks Zehra Hoca for contributing your time and for everything.

I am also very grateful to my committee members, Professor Barış KORKMAZ and Professor Hatice KÖSE for providing many suggestions and valuable discussions that improved this thesis. My sincere thanks also go to Professor Binyamin BİRKAN, who guided us in selection of experimental design. I appreciate his willingness to share his expertise knowledge in the area of autism. I am also thankful to my friend Mr. Fadi HANIA for his help on the programming of the web application.

We would like to thank the educators at the Right to Live Society (RTL) and Palestinian Society for Autism & Rehabilitation (PSAR) for their help conducting the experiments. This study would not have happened without children who participated in this study. I respect their efforts and thank their family for their cooperation.

Finally, a very special thanks to my wife Azhar and my five children. Without their love and support, this work would not have been completed. I would like to dedicate this thesis to my parents for their boundless love.

June 2016 Akram M. RADWAN

TABLE OF CONTENTS

Page

FOREWORD ..... **ix**
TABLE OF CONTENTS ..... **xi**
ABBREVIATIONS ..... **xiii**
SYMBOLS ..... **xv**
LIST OF TABLES ..... **xvii**
LIST OF FIGURES ..... **xix**
SUMMARY ..... **xxi**
ÖZET ..... **xxiii**
1. INTRODUCTION **1**
1.1 Autism and Learning Disability 1
1.2 Machine Learning for Autism 2
1.3 Purpose of Thesis 4
1.4 Thesis Research Questions 4
1.5 Thesis Contributions 5
1.6 Publications and Derived Work 6
1.7 Thesis Organization 6
2. ACTIVE MACHINE LEARNING FRAMEWORK FOR TEACHING **9**
2.1 Active Machine Learning 9
2.2 Proposed AML Framework 11
2.2.1 Formulation 11
2.3 Method 14
2.3.1 Participants 14
2.3.2 Ethical approval 16
2.3.3 Setting 16
2.3.4 Target objects 17
2.3.5 The web application 18
2.3.6 System architecture 19
2.3.7 Experimental design 20
2.3.8 Dependent measures 21
2.3.9 Procedures 21
2.3.10 Inter-observer agreement and procedural ntegrity 25
2.4 Results 26
2.4.1 Ali 27
2.4.2 Osman 29
2.4.3 Hamid 30
2.4.4 Layla 31
2.4.5 Khalil 33
2.4.6 Reduction in required teaching trials 33
2.5 Discussion 34
2.5.1 The suitability of AML framework for students with ASD 35
2.5.2 Efficiency of the teaching condition 35
2.5.3 Limitations and future directions 36

3. OBJECT CATEGORIZATION TASK FOR STUDENTS WITH ASD 39
3.1 Using Assistive Technology to Enhance Teaching 39
3.2 Results and Discussion 40
3.2.1 General performance of students 40
3.2.2 Relationships between variables 40
3.2.3 The effectiveness of difficulty levels 41
3.2.4 The effect of the negative responses 42
3.2.5 The effect of the location of object's image 43
3.2.6 The familiarity of categories 44
3.2.7 Difficulty of recognizing the objects 45
3.2.8 Guidance for developing tablet applications 47
4. CLASS IMBALANCE AND CLASS NOISE 49
4.1 Background 49
4.2 Class Imbalance Approaches 51
4.2.1 Data re-sampling methods 51
4.2.1.1 Random over-sampling 51
4.2.1.2 Random under-sampling 52
4.2.1.3 SMOTE 52
4.2.2 Cost-sensitive learning 52
4.2.3 Thresholding (cut-off adjustment) 53
4.3 Evaluation Metrics in Imbalanced Domains 53
4.4 Noise Detection Approaches 55
4.5 Machine Learning for predicting student performance 55
4.6 Literature Review 56
5. LEARNING PERFORMANCE PREDICTION ON EDUCATION DATA .. 59
5.1 Proposed Methods 60
5.2 Dataset 62
5.3 Experiments and Results 64
5.3.1 Experiment 1 64
5.3.2 Experiment 2 65
5.3.3 Experiment 3 72
5.4 Important Features on Predicting Student’s Performance 74
6. CONCLUSIONS AND FUTURE WORK 77
REFERENCES 81
APPENDICES 89
APPENDIX A 90
APPENDIX B 91
APPENDIX C 95
CURRICULUM VITAE 97

ABBREVIATIONS

ABA : Applied Behavior Analysis
AI : Artificial intelligence
AML : Active Machine Learning
ASD : Autism Spectrum Disorders
ATP : Alternating Treatments Phase
AUC : Area Under the ROC Curve
BST-CF : class-Balanced by SMOTE & Thresholding combined with CF
BST-EF : class-Balanced by SMOTE & Thresholding combined with EF
CARS : Childhood Autism Rating Scale
CF : Classification Filter
DS : Dataset
DTT : Discrete Trial Teaching
EDM : Educational Data Mining
EF : Ensemble Filter
IES : Inverse Efficiency Score
IPF : Iterative-Partitioning Filter
ITS : Intelligent Tutoring Systems
LR : Logistic Regrssion
ML : Machine Learning
PIQ : Performance Intelligent Quotes
PL : Passive Learning
PSAR : Palestinian Society for Autism & Rehabilitation
RF : Random Forest
RN : Random Classifier
ROC : Receiver Operating Characteristics
ROS : Random over-sampling
RTL : Right to Live Society
SMOTE : Synthetic Minority Oversampling Technique
SMOTE-ENN: SMOTE with Edited Nearest Neighbor
SMOTE-TL : SMOTE with Tomek links
SSL : Semi-Supervised Learning
SVM : Support Vector Machine
WISC : Wechsler Intelligence Scale for Children
WRAML : Weighted Response Active Machine Learning

SYMBOLS

α	**:** Significance level
$\boldsymbol{E_w}$	**:** Weighted error
$\boldsymbol{S_j^{(t)}}$	**:** Teaching set
$\boldsymbol{U_j}$	**:** Uncertainty measurement
$\boldsymbol{m_j^{(t)}}$	**:** Number of mistake at level L_j
$\boldsymbol{\theta^*}$	**:** Threshold value
AR	**:** Auditory repeats
$\boldsymbol{H}$	**:** Entropy
$\boldsymbol{L}$	**:** Labeled set
$\boldsymbol{L_j}$	**:** Difficulty level j
RC	**:** Response correct
RL	**:** Response latency
RT	**:** Response time
$\boldsymbol{w_j}$	: Weight for level L_j
$\boldsymbol{Im(f)}$	**:** Importance of feature
$\boldsymbol{\eta}$	**:** Learning rate
U	**:** Unlabeled set

LIST OF TABLES

Page

Table 2.1 : The performance of participants.. ..... **9**
Table 2.2 : The overall performance of categories..... **15**
Table 2.3 : Mean performance for all levels across baseline and ATP phases **28**
Table 2.4 : Average accuracy, mean of correct response latency and the IES for treatments PL and AML during ATP phase **29**
Table 3.1 : The performance of participants.. ..... **40**
Table 3.2 : The overall performance of categories..... **46**
Table 4.1 : Confusion matrix for two-class problem **53**
Table 5.1 : Description of the dataset used **63**
Table 5.2 : The results of classification algorithms using baseline feature set.. ..... **65**
Table 5.3 : Comparison of AUC results for different classifiers using baseline feature set..... **67**
Table 5.4 : Comparison of AUC results for different classifiers using baseline with RT features..... **67**
Table 5.5 : Results for BST-CF with different borderline's width on dataset DS.... **68**
Table 5.6 : Results for BST-EF with different borderline's width on dataset DS.. .. **68**
Table 5.7 : Results for BST-CF with different borderline's width on dataset DS-2..**68**
Table 5.8 : Results for BST-FF with different borderline's width on dataset DS-2..**69**
Table 5.9 : The Friedman test **72**
Table 5.10 : The AUC results based on SMOTE-ENN and weighting..... **73**
Table 5.11 : The AUC results for sequential prediction based on SMOTE-ENN & using proposed methods.. ..... **73**
Table 5.12 : The AUC results for sequential prediction based on Weighting and using proposed methods.. ..... **73**
Table 5.13 : Importance of attributes using RF model..... **74**
Table A.1 : List of target categories and objects..... **90**

LIST OF FIGURES

Page

Figure 2.1 : Participant's activities during the conduction of the experiments. **17**
Figure 2.2 : A sample of images from the dataset used for teaching and assessment. Columns L1-L4 show different difficulty levels. **18**
Figure 2.3 : (a) Screen display of baseline, teaching and assessment trial (b) The reward screen for the correct answer in a teaching trial. (c) Screen display of an initial assessment trial. **19**
Figure 2.4 : System architecture of the web application. **20**
Figure 2.5 : Flowchart of an experiment for a participant **22**
Figure 2.6 : Flowchart of the assessment session in baseline phase. **23**
Figure 2.7 : Sample of outline drawing images **24**
Figure 2.8 : Session schedule of intervention for PL and AML treatments **25**
Figure 2.9 : Screens display of AML session **25**
Figure 2.10 : A sample of details result of an assessment session **26**
Figure 2.11 : Accuracy during baseline, ATP and final phases for Ali **27**
Figure 2.12 : Mean RL during baseline, ATP and final phases for Ali **29**
Figure 2.13 : Accuracy during baseline, ATP and final phases for Osman **30**
Figure 2.14 : Mean RL during baseline, ATP and final phases for Osman **30**
Figure 2.15 : Accuracy during baseline, ATP and final phases for Hamid **31**
Figure 2.16 : Mean RL during baseline, ATP and final phases for Hamid **31**
Figure 2.17 : Accuracy during baseline, ATP and final phases for Layla **32**
Figure 2.18 : Mean RL during baseline, ATP and final phases for Layla **32**
Figure 2.19 : Accuracy during baseline, ATP and final phases for Khalil **33**
Figure 2.20 : Mean RL during baseline, ATP and final phases for Khalil **34**
Figure 2.21 : Mean accuracy at only included levels in all phases **36**
Figure 3.1 : The performance of the participants in terms of accuracy and RT **41**
Figure 3.2 : Residuals plots against the model fit the data. **41**
Figure 3.3 : The performance of the difficulty levels in terms of accuracy and RT. **42**
Figure 3.4 : The mean RT for correct and incorrect response trials. **43**
Figure 3.5 : (Left) Chosen location vs. number of errors (Right) Chosen location vs. percentage of auditory repeats. **44**
Figure 3.6 : Confusion matrix for each category **45**
Figure 3.7 : Objects vs. inverse efficiency score (IES) **46**
Figure 4.1 : The three types of instances: safe, borderline and noisy instances. **50**
Figure 5.1 : The overall design of BST-CF and BST-EF.. ..... **62**
Figure 5.2 : Average AUC obtained by different classifiers using only class-imbalance approaches and with BST-CF on dataset DS. **70**
Figure 5.3 : Average AUC obtained by different classifiers using only class-imbalance approaches and with BST-CF on dataset DS-2 ..... **70**
Figure 5.4 : Average AUC obtained by different classifiers using only class-imbalance approaches and with BST-EF on dataset DS. **71**
Figure 5.5 : Average AUC obtained by different classifiers using only class-imbalance approaches and with BST-EF on dataset DS-2 ..... **71**

Figure 5.6 : Important of all features based on RF model **75**
Figure B.1 : Ethical committee approval **91**
Figure B.2: Consent form for parents **92**
Figure B.3: Permission from RTL School **93**
Figure B.4:. Permission of PSAR School **94**
Figure C.1:. The initial probability list for participant 1..... **95**

USING MACHINE LEARNING TECHNIQUES TO ENHANCE TEACHING AND PERFORMANCE PREDICTION OF STUDENTS WITH AUTISM SPECTRUM DISORDERS

SUMMARY

The number of children diagnosed with autism spectrum disorders (ASD) is on the rise. Thanks to the advances in IT technology the use of computerized teaching systems and hence the variety, quality and quantity of data collected with them has increased. This thesis introduces a web- based teaching tool tailored for ASD students. Then, machine learning techniques, such as active machine learning (AML), noise elimination and class balancing, are applied on the collected student learning data. Using a new AML method, a method of planning what will be taught at the next session based on previous session performance is introduced. Using a new noise elimination and class balancing method to predict student performance during teaching is also devised.

The thesis represents the first attempt to use machine learning techniques for teaching students with ASD. AML technique enables a machine learning model to perform better with less labeled training data by allowing it to choose the training instances. Although AML has been studied in different domains, such as video annotation and web page classification, its applications to human learning have been studied very little. To the best of our knowledge, there is no research on using AML for teaching children with ASD.

In the first part of this thesis, we propose an AML approach for teaching students with ASD, and compare the effectiveness of passive learning (PL) and AML on teaching object recognition for those students. AML approach presents to the child the most informative teaching set of objects. For this purpose, a web and touch-based application was developed and presented on a tablet PC. Objects from everyday lives of children were grouped based on their categories and difficulty levels. The teaching procedure was based on applied behavioral analysis (ABA) principles. The study was approved by the Ethics Board of Istanbul Technical University (ITU). Five children with mild to moderate levels of ASD participated in the experiment. An alternating treatments design of single subject research methods was used to compare the effects of AML and PL. The results indicate that AML was more effective than PL for four out of the five students. Consequently, children can learn faster and are able to reach a learning criterion with fewer teaching trials.

The second part of the thesis presents an analysis and visualization of the data collected using the web application during the learning sessions. The results of the analysis were interpreted and they provided us a better understanding about object categorization for students with ASD. A number of factors influence a student's response including familiarity of categories, difficulty levels, the location of object's image and similarities with the other objects. The findings help us enhance the teaching

process and monitor student's learning progress and then personalize learning for each student. Due to several features for tablet PC, we suggest that the tablet is an effective educational tool to enhance teaching for the students. Finally, we offer suggestions that may help in future applications that can be developed for students with ASD.

In the third part of this thesis, we propose to apply machine learning techniques to predict the performance of students with ASD while they learn tasks described above. The real dataset, gathered from the web application for the students, suffers from two major problems that can negatively impact the ability of classification models to predict the correct label: class imbalance and class noise.

A series of experiments have been carried out to improve prediction results through reducing the effect of both class noise and class imbalance on the induced classifiers. In the first experiment different approaches have been applied in order to resolve the problem of imbalanced classification by data re-sampling methods and cost-sensitive learning. In the second experiment, two methods were proposed to eliminate the noisy instances, located inside the borderline area, from the majority class. Our methods combine over-sampling SMOTE technique with thresholding technique to balance the training data and choose the best boundary between classes. Then we apply a noise detection approach to identify noisy instances. The best results from the two experiments were used in the third experiment to predict the response correct of the students' future trial using only the preceding session's trials. Results for different classification algorithms showed that the two proposed methods achieve significant improvements in terms of the AUC (Area Under the ROC Curve) metric with respect to the existing techniques used for class-imbalance problem. Moreover, our prediction methods performed better when data of a student, who has a lot of noise, were excluded from the dataset. The most robust classifier over our methods was random forest. It performed better on average than logistic regression (RL), Support vector machine (SVM) and Adaboost with LR.

The contributions of this thesis are as follows:

1- At the time of writing, this is the first study of using AML to teach students with ASD.
2- The research results enhanced education outcomes and student's recognition abilities, and thus can improve the learning process of a student with ASD to reach better results.
3- The thesis provides a dataset of a real-world images for the most common categories of concrete objects in daily life of a child.
4- We develop and implement a user-friendly web-based application, which presents a customizable and efficient tool enabling the child, parents and caregivers to work together through the computerized learning methods.
5- We propose to use machine learning filtering techniques on students' learning data to improve the quality of the data.
6- The prediction results have significantly improved when our proposed methods are combined with class-imbalance approaches.

YAPAY ÖĞRENME YÖNTEMLERİ İLE OTİZM SPEKTRUM BOZUKLUĞU OLAN ÖĞRENCİLERİN ÖĞRETİMİNİN VE ÖĞRETİM PERFORMANSI TAHMİNİNİN İYİLEŞTİRİLMESİ

ÖZET

Otizm Spektrum Bozukluğu (ASD – Autism Spectrum Disorder) teşhisi alan çocuk sayısı son zamanlarda büyük artış göstermiştir. Internet Teknolojileri'ndeki ilerlemeler sayesinde ise bilgisayarla öğretim sistemlerinin kullanımı ve dolayısı ile bu yolla toplanan verilerin çeşitliliği, miktarı ve kalitesi artmıştır. Bu tez kapsamında, öncelikle ASD'li çocukların öğretiminde kullanılabilecek web tabanlı bir yazılım üretilmiştir. Daha sonra öğrencilerle yapılan öğrenim uygulamalarından elde edilen veriler üzerinde aktif yapay öğrenme, gürültülü verileri giderme, sınıfları dengeleme yapay öğrenme teknikleri kullanılmıştır. Tez kapsamında geliştirilen yeni bir aktif yapay öğrenme tekniği ile önceki öğrenme oturumlarının verileri kullanılarak bir sonraki oturumda ne öğretilebileceği belirlenmiştir. Yine tez kapsamında geliştirilen ve gürültü temizleme ve sınıf dengelemeyi aynı zamanda yapan bir yöntemle, öğretim sırasında öğrenci başarım ve süre performansını tahmin eden bir yöntem geliştirilmiştir.

Bu tez otizm spektrum bozuklukları (ASD – autism spectrum disorders) olan çocukların eğitiminde yapay öğrenme (ML – machine learning) teknikleri kullanılarak çocuklara bir nesne tanıma görevi öğretilirken performanslarının öngörülmesine yönelik ilk girişimi yansıtmaktadır. Otizm genelde öğrenme bozukluğuyla ilişkilendirilir. ASD'li insanların yaklaşık %70'inde belli bir dereceye kadar öğrenme bozukluğu mevcuttur. Dolayısıyla bu kişiler çoğu insanla aynı şekilde öğrenemez ve bir kavramı öğrenmeleri için eğitimlerinde özel yöntemlerin izlenmesine ihtiyaç vardır.

Bu tezde davranış modelleri ile ML modelleri birlikte kullanılarak bu çocukların daha iyi öğrenmeleri için ML tekniklerinden faydalanmalarının mümkün olup olmadığı araştırılmıştır. Aktif makine öğrenimi (AML – Active Machine Learning) teknikleri, ML modelinin öğrenme sırasında kullanılacak örnekleri seçmesine olanak tanıyarak modelin daha az sayıda etiketli eğitim verisi ile daha iyi işlev görmesini sağlamaktadır. AML, video etiketleme ve internet sayfası sınıflandırması gibi farklı alanlarda kullanılmış olsa da, insan öğrenmesi üzerindeki uygulamalarına yönelik araştırmalar kısıtlıdır. Bildiğimiz kadarı ile ASD'li çocuklara eğitim vermek için AML tekniklerinin kullanılması üzerine bir araştırma mevcut değildir.

Bu tezin ilk kısmında ASD'li çocuklara eğitim verilmesi konusunda bir AML yaklaşımı benimsenmesi önerilmiş ve bu çocuklar için nesne tanınmasının öğretilmesinde pasif öğrenme (PL - Passive Learning) ve AML tekniklerinin etkililikleri karşılaştırılmıştır. AML yaklaşımı çocuklara bir dizi nesnenin öğretilmesi için, çocuğun daha önceki eğitim seanslarında aynı kategoride ya da zorluk seviyesinde olan nesneler tanıma sırasında gösterdiği performansı hesaba katan bir

yöntem sunmaktadır. Önceki eğitimler sırasındaki performans, çocuğun bilip bilmediği, ne kadar sürede cevap verdiği ve kaç defa yönergenin sesli olarak tekrar edilmesini istediği bilgilerinin hepsini içerecek şekilde değerlendirilmiştir. Bu doğrultuda internet ve dokunma tabanlı bir uygulama geliştirilmiş ve bir tablet bilgisayar üzerinde sunulmuştur. Çocukların günlük hayatlarından nesneler kategorilerine ve dört zorluk derecesine göre gruplara ayrılmıştır. Öğretme prosedürü için ASD'li her yaştan insan için etkili olduğu bulunan uygulamalı davranış analizi (ABA – Applied Behavioral Analysis) prensipleri temel alınmıştır. Uygulama çocuğun hata yapmasına izin vermemektedir. Eğer çocuk doğru seçeneği bulamamış ise, sunulan resimler üzerinde çeşitli etkiler ile (yanlış resmin küçültülmesi, doğru resmin sallanması gibi) çocuk doğru seçeneğe yönlendirilmektedir. Çocuğa bulması gereken nesne sesli olarak bildirilmektedir. Doğru ya da yanlış cevap vermesi durumunda, doğru cevabı pekiştirecek ve doğruyu ödüllendirecek sesli uyarılar içermektedir. Araştırmaya hafif ve orta derecede ASD'si olan beş çocuk (1 kız, 4 erkek) katılmıştır. Katılımcıların yaşları 5 ile 9 arasındadır (ortalama: 7,16; standart sapma: 1,1). Katılımcılar Filistin'de ASD'li öğrencilere eğitim veren iki okuldan seçilmiştir. AML ve PL'nin etkilerinin karşılaştırılması için tek denekli araştırma yöntemleri şeklinde tasarlanan eğitimler değişmeli olarak uygulanmıştır. Çalışma İstanbul Teknik Üniversitesi (İTÜ) Etik Kurulu tarafından onaylanmıştır.

Sonuçlar AML'nin araştırmamızdaki beş öğrencinden dördü için pasif öğrenmeden daha etkili olduğunu göstermiştir. Sonuç olarak bu öğrenciler bir öğrenme kriterine ulaşmak için gerekli öğrenme denemelerine daha az ihtiyaç duyarak daha hızlı öğrenebilmiştir. Bu tezde tanımlandığı şekliyle önerilen AML sadece ASD'li bireylerin eğitiminde kullanılmakla sınırlı değildir, başka eğitim alanlarına da uygulanabilir.

Tablet bilgisayarlar, özellikle anında geri bildirim sağlamaları ve öğrenme sürecindeki ara adımlar hakkında geniş bilgi tutma imkanları sayesinde öğretimin geliştirilmesi açısından etkili eğitim araçları olabilmektedir. Tezin ikinci kısmında öğrenme seansları esnasında, tez kapsamında geliştirilmiş olan bir web tabanlı internet uygulaması tablet bilgisayarlar üzerinde kullanılarak, toplanan veriler analiz edilmiş ve görselleştirilmiştir. Analiz sonuçları sayesinde deney yapılan ASD'li öğrencilerin nesne sınıflandırmasına ilişkin açıklamalar sunulmuştur. Kategorilerin aşinalığı, zorluk seviyeleri, nesnenin görüntüsünün konumu ve diğer nesnelerle olan benzerliği gibi birçok faktör öğrencinin tepkisini etkileyebilir. Çalışmamızda tepki süresi ile öğrencilerin performansları arasında güçlü bir negatif korelasyon olduğu saptanmıştır. Aynı zamanda pozitif tepkilerin negatif tepkilerden daha hızlı verildiği gözlemlenmiştir. Bu bulgular ile öğrencilerin öğrenme süreçleri izlenerek öğrenmenin her bir öğrenci için kişiselleştirilmesi sağlanabilecektir. Son olarak da ASD'li öğrenciler için ileride geliştirilebilecek uygulamalara yardımcı olabilecek önerilerde bulunulmaktadır.

Tezin üçüncü kısmında ASD'li öğrenciler yukarıdaki nesne tanıma görevini öğrenirken performanslarının öngörülmesi için makine öğrenimi tekniklerinin uygulanması tavsiye edilmektedir. Öğrenciler için geliştirilmiş olan internet uygulamasından elde edilen gerçek veri kümesinde doğru etiketi (öğrencinin bir sonraki soruyu bilip bilmeyeceği ve ne kadar sürede bileceği) tahmin etmek için sınıflandırma modellerinin yeterliliğini olumsuz bir şekilde etkileyebilecek iki büyük problemin olduğu görülmüştür: sınıf dengesizliği ve sınıf gürültüsü.

Hem sınıf gürültüsünün hem de sınıf dengesizliğinin uyarılan sınıflandırıcılar üzerindeki etkisinin azaltılması ile öngörü sonuçlarının iyileştirilmesi için yapılan farklı araştırmalar vardır. Dengesiz sınıflandırma problemini çözmek amacıyla verilerin tekrar örneklenmesi ve maliyete duyarlı öğrenme yaklaşımları uygulanmıştır. Sınıfın çoğunluğuna göre sınır bölgesi içerisinde bulunan gürültülü durumların ortadan kaldırılması için iki yöntem tavsiye edilmiştir. Tez kapsamında geliştirilen yöntemlerde, verilerde sentetik azınlık aşırı örneklendirme tekniği olan SMOTE (synthetic minority over-sampling technique) ile eşik tekniği kullanılarak eğitim verilerinin dengelenmesi ve sınıflar arasındaki en iyi sınırın seçilmesi sağlanmaktadır. Daha sonra gürültülü durumların belirlenmesi için bir gürültü tespit yaklaşımı uygulanmıştır. İlk yöntemde gürültülü durumların belirlenmesi için sınıflandırma filtresi (classification filter - CF) kullanılırken, ikinci yöntemde toplu filtre (ensemble filter - EF) kullanılmıştır. İki deneydeki en iyi sonuçlar üçüncü deneyde yalnızca önceki seanstaki deneyler ile öğrencilerin ilerideki deneydeki doğru yanıtlarının öngörülmesi için kullanılmıştır.

Farklı sınıflandırma algoritmalarının sonuçları önerilen CF ve EF bazlı iki yöntemin daha önceki çalışmalarda kullanılan sınıf dengesizliği tekniklerine göre başarım açısından önemli ilerlemeler sağladığını göstermektedir. Ayrıca verisinde çok sayıda gürültülü örnek olan bir öğrencinin verileri veri kümesinden tamamen çıkarıldığında, sınıflandırma yöntemleri daha da etkili olmuştur. Farklı sınıflandırıcılar kullanılan deneylerde, en iyi sınıflandırıcı rastgele orman sınıflandırıcısı olmuştur. Rastgele orman sınıflandırıcısı ortalama olarak lojistik regresyon (LR), destek vektör makineleri (DVM) ve LR sınıflandırıcısı kullanan Adaboost yönteminden daha etkili sonuçlar vermiştir. Özelliklerin önemi üzerine yapılan bir analiz, tepki süresi ve nesnenin seviyesinin bir öğrencinin performans tahmininin doğru olmasında en etkili olduğunu göstermiştir.

Bu tezin katkıları şu şekilde özetlenebilir:

1- Bu yazımın yapıldığı sırada, ASD'de olan öğrencilerin eğitiminde aktif yapay öğrenme (AML) tekniklerinin kullanıldığı ilk çalışmadır.
2- Yapılan tez çalışması sonucunda oluşturulan yöntemlerle öğrencilerin bilgi öğrenme performanslarının diğer yöntemlere göre daha iyi arttığı gösterilmiştir. Bu yöntemler ASD'li öğrencilerin öğrenme süreçlerini iyileştirme amacı ile kullanılabilir.
3- Tez çalışması sırasında bir çocuğun günlük hayatında gördüğü nesnelerin değişik zorluk seviyelerine göre sınıflandırıldığı bir görüntü veri kümesi oluşturulmuştur.
4- Kullanımı kolay web tabanlı bir uygulama tasarlanmış ve üretilmiştir. Bu uygulama öğrenci, ebeveyn ve öğretmenin beraber çalışarak bilgisayarla öğretim yöntemlerini kullanabileceği bir ortam oluşturmuştur.
5- Geliştirilen verideki gürültülü örnekleri bulup eleyen filtreleme yöntemlerinin veri kalitesini ve başarım performansını arttırdığı gösterilmiştir.
6- Gürültü giderme yöntemleri sınıf dengesizliğini giderme yöntemleri ile bir arada kullanılarak başarım performansı daha da arttırılmıştır.

1. INTRODUCTION

This chapter first introduces the area of autism and some difficulties of learning disability associated with autism. Second, it places our research presented in this thesis into a broader machine learning context. Next, it presents the purpose of the thesis and identifies the research problems, states the research questions, and lists the thesis contributions. Finally, the structure of the rest of this thesis is outlined.

1.1 Autism and Learning Disability

Autism spectrum disorder (ASD) refers to a group of complex neurodevelopment disorders. ASD is characterized by significant impairment in some areas of functioning, including social interaction, verbal and nonverbal communication, social imagination and repetitive behaviors (American Psychiatric Association, 2013). An estimated over 3 million individuals in the U.S. and tens of millions worldwide are affected by ASD (Autism Speaks, 2015). There is no single medical test for diagnosis of autism and no known cure so far.

Autism is often associated with learning disability. Nearly 70% of people with ASD have some degree of learning disability (Corbett, 2006) based on the individual's IQ score and the level of cognitive functioning (Jordan, 2013). Hence, students with ASD cannot learn in the same way as most people, and they need special treatment to learn a concept or an object. For every concept, there is a need to repetitively teach it using different instances and after a carefully chosen regime. In the human learning scenario, both obtaining training instances and predictions produced by the child for instances may be difficult. During the last decades, there have been efforts to understand the nature of the learning difficulties in ASD and to find effective treatments. One of the difficulties faced by people with ASD is the recognition of categories.

Categorization of objects poses challenges to children with ASD, especially when it is based on complex and less apparent features such as material or contextual features (Tager-Flusberg, 1985; Klinger and Dawson, 1995; Galleguillos and Belongie, 2010).

People with ASD also have deficiencies in prototype and category formation (Gastgeb et al, 2012). Plaisted et al. (1998) concluded that individuals with ASD were more accurate in processing unique features of objects than other commonly shared features between objects. Gastgeb et al. (2006) examined whether individuals with ASD have difficulty categorizing common objects, and whether their categorization abilities are affected by the typicality of objects. The authors found that "Individuals with ASD can readily categorize when the task involves simple and typical basic objects or faces but have difficulty when categorization is more complex or involves less typical objects or faces" (Gastgeb et al, 2012). .

There is no single teaching strategy that will be successful with all children with ASD. One of the methods that has shown effectiveness in teaching skills to children with ASD is Discrete Trial Teaching (DTT) (Artoni et al, 2012; Eldevik et al, 2013). DTT is an instructional approach which divides teaching skill or concept into parts/elementary steps with increasing levels of difficulty and teach repetitively to accomplish the task successfully. DTT is a specific method of teaching based on the principles of Applied Behavior Analysis or ABA (Alberto and Troutman, 2012). The procedures based on ABA have been shown to be efficient for people of any age with ASD in a number of research studies (Artoni et al, 2012; NAC, 2015). ABA aims to help individuals to give correct answers. Prompting the individual to perform the desired skill is an important part of this method (Alberto and Troutman, 2012). These prompts can be provided as visual cues, such as pointing out the correct answer or fading out the wrong options. Prompts need to be repeated until the desired behavior or skill is performed. Reinforcement lets the child be rewarded for the correct behavior, and it should immediately follow the correct response to make the student understand the connection between the correct answer and the reward (Doenyas et al, 2014).

Educational therapy techniques, such as ABA, Treatment and Education of Autistic and related Communication handicapped CHildren (TEACCH) (Mesibov and Shea, 2010) and Assistive Technology (Lang et al, 2014) without the use of any machine learning algorithms have been used in order to teach children with ASD.

1.2 Machine Learning for Autism

With the growing amount of data produced daily, the need of techniques to handle these data has been increased. Machine learning (ML) is a prominent area of computer

science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence (AI). ML focus on analyzing data to identify patterns and make accurate predictions. Over the past decade, learning algorithms have found widespread applications in numerous areas such as computer vision, object recognition, web search, natural language processing, emotion recognition etc.

Supervised learning (also called inductive learning) learns from the available data (experience), which is given in the form of training data (instances). The knowledge induced from the data can then be used for descriptive or predictive purposes. Supervised ML methods have been employed in ASD studies in order to enhance autism diagnosis (Wall et al, 2012; Bone et al, 2015) and to discriminate children with ASD from typically developing children (Crippa et al, 2015). Lastly, several ML classifiers have been applied to test whether a subset of behaviors was sufficient to detect individuals with and without ASD with high accuracy (Kosmicki et al, 2015). There are a few studies based on reinforcement learning (RL) (Sarma and Ravindran, 2007; Solomon et al, 2011; Wang, 2014). RL (Sutton and Barto, 1998) is an area of ML inspired by behaviorist psychology. It allows machines and software agents to automatically determine the ideal behavior or action within a specific context, in order to maximize its performance. Sarma and Ravindran (2007) proposed to use RL for building intelligent tutoring systems (ITS) to teach students with ASD, who cannot communicate well with others. ITS used neural networks to customize instructions according to the student's needs. The study by Solomon et al. (2011) investigated the probabilistic reinforcement learning with Bayesian state–space model in adults with ASD. They showed that individuals with ASD have further deficits in using positive feedback to exploit rewarded choices in a task requiring learning relationships between stimulus pairs.

Semi-supervised learning (or SSL) has attracted a highly considerable amount of interest in ML. SSL techniques allow classifiers or learners to learn from labeled and unlabeled data at the same time (Davy, 2005; Zhu and Goldberg, 2009). Typically, they are used when we have a small size labeled dataset with a large size unlabeled dataset. During the last decade, SSL methods such as active learning, co-training, and co-testing have significantly improved learning performance in various applications. Actually human learns concepts in similar way of SSL; from a limited number of labeled data (e.g. parental labeling of objects) and a large amount of unlabeled data

(e.g. observation of objects without naming in real-life) (Zhu and Goldberg, 2009). In the ML scenario, it is easy to obtain the predictions of the classifier and it is usually expensive to obtain the actual labels for instances.

Our research is motivated by the fact "As the psychological and machine learning models are formally identical, the lifted variants of the machine learning models provide candidate hypotheses about human SSL" (Gibson et al, 2013, p. 135). Because humans use both labeled and unlabeled data when learning a categorization task (Zhu et al, 2007), SSL techniques can be useful for explaining human behavior.

1.3 Purpose of Thesis

In this thesis, we aim to incorporate ML models with behavioral models to teach students with ASD and to investigate whether they can benefit from ML techniques to learn better. The first goal is to examine the useo of the active machine learning (AML) technique for teaching students with ASD to recognize object categories, and evaluate its effectiveness compared with passive learning (PL). Secondly, we aim to use ML filtering techniques, on imbalanced data gathered from the web application during learning sessions, to improve the quality of the data. Then we can get more accurate prediction results of the student’s performance through reducing the impact of both class noise and class imbalance on the induced classifiers. For this purpose, two empirical methods are proposed to address noise filtering and class imbalance problems simultaneously.

1.4 Thesis Research Questions

To handle the research problems, we address the following research questions:

Q1: Which technique PL or AML is more effective on teaching object recognition for students with ASD?

Q2: Can AML reduce the number of teaching trials required to reach a learning criterion?

Q3: Do the object difficulty levels have an effect on object categorization?

Q4: Which categories of objects are the most familiar to children with ASD?

Q5: Which factors have more influence on a student's response?

Q6: How we can reduce the impact of both class noise and class imbalance problems on the students' learning data?

Q7: Are the AUC scores obtained by the combinations of proposed methods in Chapter 5 and a classifier significantly better than the AUC scores obtained from only a classifier with class-imbalance approach?

Q8: Which attributes and features are the most significant predictors of the correctness of a student?

The thesis aims to answer these research questions by applying our investigation on a group of students with ASD. Answers to questions Q1 and Q2 given in Chapter 2, where question Q3-Q5 are examined in quantitative analyses of gathered data in Chapter 3. Answers to questions Q6-Q8 are found by results of a series of experiments in Chapter 5.

1.5 Thesis Contributions

The contributions of this thesis are as follows:

1- At the time of writing, this is the first study of using AML to teach students with ASD.

2- The research results enhanced education outcomes and student's recognition abilities, and thus can improve the learning process of a student with ASD to reach better results.

3- The thesis provides a dataset of a real-world images for the most common categories of concrete objects in daily life of a child.

4- We develop and implement a user-friendly web-based application, which presents a customizable and efficient tool enabling the child, parents and caregivers to work together through the computerized learning methods.

5- We propose to use machine learning filtering techniques on students' learning data to improve the quality of the data.

6- The prediction results have significantly improved when our proposed methods are combined with class-imbalance approaches.

1.6 Publications and Derived Work

The scientific contributions of the thesis are listed in the following articles:

- **Radwan, A.**, and Cataltepe, Z., (2016). The Use of Tablet PCs in Teaching Object Recognition to Students with ASD, *Proceedings of 3rd International Conference on Education and Social Sciences INTCESS*, February 8-10, 2016, Istanbul, Turkey, pp. 399-408.
- **Radwan, A.**, and Cataltepe, Z., (2016). Using assistive technology to enhance teaching for students with autism spectrum disorders. *International E-Journal of Advances in Education*, *2*(4), 112–121. DOI: http://dx.doi.org/10.18768/ijaedu.15760
- **Radwan, A. M.**, Birkan, B., Hania, F., and Cataltepe, Z., (2016). Active machine learning framework for teaching object recognition skills to children with autism. *International Journal of Developmental Disabilities*. Published online 17 Jun 2016. DOI:10.1080/20473869.2016.1190543
- **Radwan, A.**, and Cataltepe Z., Improving Performance Prediction on Education Data with Noise and Class Imbalance. *Intelligent automation and Soft Computing*. Submitted, July 2016.

1.7 Thesis Organization

The remainder of the thesis is organized as follows. Chapter 2 presents the proposed active machine learning framework, which teaches object recognition to the student with ASD by presenting to him the most informative teaching set of objects. The effectiveness of framework is compared to the traditional passive learning. Additionally, the design of web application and the procedure of assessment and teaching trials are described.

Chapter 3 presents an analysis and visualization of the data collected using our web application during the learning sessions and concludes some useful decisions about teaching process and student's response.

Chapter 4 describes several approaches that have been used to handle the class imbalance and noise detection. Furthermore, we review some previous works related to predicting student's performance using ML methods.

Chapter 5 proposes two empirical methods that address noise filtering and class imbalance problems simultaneously. It also describes three series of experiments and the important of features used in the prediction.

Chapter 6 summarizes the results presented in this thesis, offers conclusions of the work performed, and presents directions for further work. Finally, Appendices provide additional materials that required to our work in Chapter 2.

2. ACTIVE MACHINE LEARNING FRAMEWORK FOR TEACHING

2.1 Active Machine Learning

When a ML model is trained, learning is performed on a random subset of all the available set of labeled training data. We will refer to this mode of learning as passive learning (PL). In PL mode of learning, the classifier (learner) does not participate interactively with the teacher (Davy, 2005). A passive learner receives a random dataset from the world and then produces a classifier or model. Thus, PL is more straightforward and easier to implement.

Active machine learning (AML) is a popular research area in ML (Davy, 2005; Settles, 2010). It allows selection of the most informative instances in training dataset of the domain for manual labeling. AML aims to produce a highly accurate classifier using as few labeled instances as possible, thereby required labeled data at minimal cost (Settles, 2010). In AML, the classifier is initially trained on a small set of instances, and it chooses informative instances from an unlabeled pool of data to request labels from the expert or oracle that can upon request provide a label for any instance in this pool (Tong and Koller, 2001). We provide a comparison of AML and TPL in Table 2.1 (Davy 2005; Settles 2010).

Table 2.1 : Comparison of active machine learning and passive learning.

PL	AML
No control over training instances	select training instances from a pool of unlabeled data (queries)
Large number of required training instances	Relatively small
Examine the entire training data before inducing a classifier (batch process)	Learner sees one or a subset instances at a time (iterative process)
One classifier induced	Many
Simple stopping criteria	Complex

Multiple studies proposed several AML algorithms and applied them to different computer recognition applications. They have shown that when AML is used, ML models require significantly less training data and can still perform well without loss of accuracy (Settles, 2010; Yang et al, 2014). Unlike machines, a human learner gets tired as s/he answers questions, so finding out whether s/he knows a concept or not (i.e. getting the labels) is usually an expensive task. Therefore, using AML for human learning may help human learning become more efficient and effective and hence reduce the cost of teaching. There are a few studies on the applications of AML to the human learning domain. The first such study was by (Zhu et al, 2007) which tried to answer the question "Do humans use unlabeled data in addition to labeled data to learn categories?". It also aimed to evaluate whether a human learner is sensitive to the distributions of unlabeled instances in classification. Results showed that human behavior fitted well to the predictions of a Gaussian mixture model for semi-supervised learning. Castro et al. (2009) investigated what they refer to as human active learning. They showed that humans learn faster with greater performance when they can actively select the informative instances from a pool of unlabeled data instead of random sampling. However, human AML is sensitive to noise, and humans were not as good as machines in selecting queries from an unlabeled dataset of artificial 3D visual stimuli. There are other empirical studies which addressed human AML with a focus on learning two-class problems on a one-dimensional input space (Zhu et al, 2007; Castro et al, 2009), but there were obstacles to generalize the model to multi-dimensional classification problems (Gibson et al, 2013).

Typically, AML approaches select a single unlabeled instance, which is the most informative at that iteration and then re-train the classifier. Even for the simplest models, the re-training process is time consuming. Thus, in this study, we selected a batch mode active learning strategy (Hoi et al, 2006; Guo and Schuurmans, 2008) that selects multiple instances each time.

There are a number of AML query selection strategies which have been presented by Settles (2010): 1) Uncertainty Sampling which is the simplest and the most commonly used strategy. Uncertainty sampling focuses on selecting the instance that the classifier is most uncertain about to label, 2) Expected Error Reduction which aims to query instance that minimizes the expected error of the classifier; and 3) Query-by-Committee (QBC) in which the most informative instance is the one that a committee

of classifiers find most disagreement. Bagging and boosting are used to generate committees of classifiers from the same dataset. They aim to combine a set of weak classifiers to create a single strong classifier. While bagging creates each base classifier independently, boosting allows these classifiers to influence each other during training process (Olsson, 2009). Boosting is an iterative process that initially assigns equal weight to each of the training samples, then the weights are modified based on the error rate of individual classifier.

2.2 Proposed AML Framework

We present a novel batch-mode pool-based AML framework (Guo and Schuurmans, 2008; Settles, 2010) on a domain containing instances with various difficulty levels to a child. The purpose of our framework is to determine how uncertain or informative an unlabeled instance is to a particular child at a certain iteration. Then we can select instances to be learned from the unlabeled pool U at the next iteration. Our approach chooses informative instances based on the commonly used uncertainty sampling strategy. This strategy selects the unlabeled instances (objects) that the child has the least confidence.

The uncertainty sampling approach is usually associated with a classifier, which is used to compute the uncertainty of each instance in an unlabeled pool U (Yang et al, 2014). AML algorithm often starts with a small size of labeled training data. After the classifier is trained, we can provide the posterior probability $P_\theta(y|x)$ under model with parameters θ. The uncertainty function can be calculated using the label probability $P_\theta(y|x)$.

In our AML framework, a child plays the role of the classifier and we do not have a probabilistic model. So, we computed the uncertainty in the context of the child's responses to measure of informativeness for all objects. If an object's uncertainty is high, it implies that the child does not have sufficient knowledge to classify the object, and then adding this object into the training set can improve the child's recognition ability.

2.2.1 Formulation

Let O_i, $i = 1, 2, .., N$ be the objects of the dataset to be learned, and L_j, $j = 1, 2, 3$ and 4 be the levels of difficulty for objects. We selected four difficulty levels

in this study, but the analysis below can be easily extended to a different number of difficulty levels. Let $L = \{(O_1, y_1), \dots, (O_m, y_m)\}$ denote the set of labeled instances where y_i is the label for object O_i. Let U denote the set of unlabeled instances. In order to measure uncertainty for each object and level, the following three factors were considered:

1. Response correct $RC(i,j)$ which is defined as follows: $RC(i,j)$ = 1 means that the object O_i is classified correctly by the child in the assessment session at level L_j and $RC(i,j)$ = 0 otherwise.
2. Response latency $RL(i,j)$ which is measured as the time the participant has taken to label an object O_i in the assessment session at level L_j.
3. The number of auditory stimulus repeats $AR(i,j)$ which is required to classify the object O_i at level L_j.

In previous works that used uncertainty for AML, classifiers' label probability was used to measure uncertainty. In this study, in addition to RC, which corresponds to label probability, we also have the other factors, which are RL and AR. Below we describe a method to combine all of these factors in uncertainty measurement. First of all, we introduce the uncertainty component that corresponds to label probability using the RC values. Since objects in our study have different difficulty levels, we weigh informativeness of an object according to its level. Thus, the weighted error of participant's response for object O_i can be defined as follows:

$$E_w(O_i) = \sum_{j=1}^{4} w_j(1 - RC(i,j)), \qquad i = 1,2,\dots,N \tag{2.1}$$

where w_j is the weight for level L_j. The weights w_j sum up to 1 and they determine how much importance should be given to each level. Especially at the beginning, the participants are more likely to make mistakes at higher levels. In order to emphasize errors at lower levels at the beginning of learning the weights should be initialized such that lower levels' weights are larger. In this study, the weights are initialized as follows:

$$w_1 = 0.4, w_2 = 0.3, w_3 = 0.2 \text{ and } w_4 = 0.1 \text{ where } \sum_{i=1}^{4} w_i = 1 \ .$$

The weights are updated after each iteration, which consists of assessment sessions at included levels in the current phase, according to the following rule:

$$w_j^{(t+1)} = w_j^{(t)}[1 - (5 - j)\eta]^{m_j^{(t)}}, \qquad j = 1,..,4 \tag{2.2}$$

where $w_j^{(t)}$ is weight j at iteration t and η is the learning rate such that $0 < \eta < {}^1/_4$. In our implementation, a small learning rate was used ($\eta = 0.05$) for all iterations. $m_j^{(t)} = \sum_{i=1}^{N} 1 - RC(i,j)$ is the number of mistakes for a participant at level L_j where $0 \leq m_j^{(t)} \leq$ N. After weights are updated, they are always renormalized so that their sum is 1. From equation (2.2), the weights for high levels are more sensitive to the number of mistakes than weights for lower levels. Thus, the increase in the value of a weight implies that the participant has improved at the corresponding level.

Entropy in information theory is associated with the orderly or disorderly configuration of data. Disorderly configuration can be interpreted as that most of the data points are scattered randomly (Li et al, 2004). We used entropy to measure the similarity (or dispersion) among the weight vector $W^{(t)} = (w_1^{(t)}, w_2^{(t)}, w_3^{(t)}, w_4^{(t)})$ at iteration t. Maximum entropy ($H_{max} = 2$) occurs when all weights are uniformly distributed across levels and $w_j^{(t)} = 0.25$ for all L_j. The decrease in entropy indicates existence of weight(s) dissimilar to other levels' weights, and this fact could be used to detect outlier weight(s) different than other weights. The above formulation of weights would enable us to determine the mastery criterion level that the participant needed for mastering or passing some levels. In current study, after performing a phase, the entropy H of weight vector is compared with the previous phase. If $H\left(W^{(t)}\right)$ decreases, then the level corresponding to a larger weight in previous iteration is assigned to be passed. This means that the participant has mastered the level, and there is no need to assess this level at the following phases.

In addition to the weighted error $E_w(O_i)$, we incorporate RL and AR into uncertainty measurement of an object O_i at level L_j as follows:

$$U_j(O_i) = E_w(O_i) \times \frac{RL(i,j)}{\sum_{i=1}^{N} RL(i,j)} \times \exp\left(\frac{AR(i,j)}{\sum_{i=1}^{N} AR(i,j)}\right)^{\beta} \tag{2.3}$$

$$i = 1,2,\ldots,N \text{ and } j = 1,..,4$$

In this equation, the effect of RL and AR has been normalized. Here β controls the importance of AR. The effect of AR has been reduced using the square ($\beta = 2$).

After each iteration, for each level L_j and based on the performance of the participant, we computed the uncertainty $U_j(O_i)$ for each object O_i at level L_j according to equation (2.3). Based on the uncertainty values, we selected a batch of n queries, which we call the teaching set $S_j^{(t)}$ at iteration t. This set contains n objects with the highest $U_j(O_i)$

values and it will be used in next phase. We summarize our framework, called the weighted response AML framework (WRAML), in Algorithm 1.

Algorithm 1 : One phase of the WRAML framework.

Given: Labeled set $\boldsymbol{L}$, unlabeled set $\boldsymbol{U}$, query batch size n,
initial values for weights $\boldsymbol{w_j^{(0)}}$, current iteration number $\boldsymbol{t}$ and weights $\boldsymbol{w_j^{(t)}}$

```
m := 1
while (m < 3) do
// In the procedure of the investigation, a phase consists of only two iterations
        m = m + 1
        t = t + 1
        Do assessment sessions at included levels
        // compute the uncertainty for each object at each level
        for level j = 1 to 4 do
            for object i = 1 to N do
                Retrieve RC(i,j), RL(i,j) and AR(i,j)
                Compute U_j(O_i) in equation (2.3)
            end for
        end for
        // compute the teaching set S_j^(t) for each level.
        for level j = 1 to 4 do
            for query k = 1 to n do
                // query the most informative instances
                Select object O_k^* in S_j^(t) with the maximum U_j(O_i)
                Query the true label y_k for O_k^*
            end for
        end for
        Update the weights w_j^(t+1) according to equation (2.2)
end while
// Decide on which levels are passed after the phase is performed
if H(W^(t)) < H(W^(t-2)) then    //t-2 because a phase consists of two iterations
        for level j = 1 to 4 do
            if w_j^(t) > w_j^(0) & w_j^(t) ≥ w_j^(t-1) then
            Assign L_j passed
        end for
end if
```

2.3 Method

2.3.1 Participants

Five children (1 female and 4 males) with mild to moderate levels of ASD participated in this study. Participants' ages ranged from 5 to 9 years (mean= 7.16, SD= 1.1) (see Table 2.2). We selected eligible participants based on the following criteria: (a) having

a diagnosis of ASD by an independent psychiatrist based on DSM-5 criteria (American Psychiatric Association, 2013); (b) having no other disabilities; (c) at least one year of full-time school enrollment; (d) attending school regularly (five days a week); and (e) being able to remain seated for 15-20 minutes while engaged in an activity. Participants were recruited from two schools in Gaza-Palestine that provide education for students with ASD; Right to Live Society (RTL) and Palestinian Society for Autism & Rehabilitation (PSAR). The Arabic version of Wechsler Intelligence Scale for Children (WISC-III) (Khaleefa, 2006) was used to assess children's IQ for determining the general level of intelligence. Moreover, the Childhood Autism Rating Scale (CARS) was used for assessing the presence and severity of ASD (Vaughan, 2011). CARS scores, which ranged from 32 to 36, indicated a moderate to mild degree of severity of autism.

Table 2.2 : Description of participants[1].

Name	Gender	Age (year)	WISC PIQ Score	CARS Score	School	Tablet Skills
Ali	M	7.3	90	34	PSAR	good
Osman	M	5.4	76	36	PSAR	poor
Hamid	M	8.8	80	33	RTL	good
Layla	F	7.5	72	36	RTL	poor
Khalil	M	6.8	70	35	RTL	average

Ali was 7 years and 4 months old boy at the beginning of the study. His performance IQ (PIQ) score on WISC-III was 90 and his score on CARS indicated a mild degree of ASD. Ali had no difficulty with receptive and expressive language. He responded quickly to the educator's instructions. He joined the PSAR School two years ago and his reading skills are good compared to his peers. His tablet usage skills were very good and also he had a high level of engagement. Therefore, he enjoyed the experiment.

Osman was 5 years and 5 months old boy, with a diagnosis of a moderate degree of autism and his PIQ was 76.. He was enrolled PSAR School 1.5 year ago. Osman presented very limited verbal skills, particularly verbal comprehension. His assessment profile indicated for self-harm and aggressive behaviors, which included kicking, hitting, and crying. These behaviors were observed both at home and at

[1] In order to protect the children's identity, fictitious gender matched names are used.

school. So we faced some problems while conducting the experiment sessions and his educator needed to be next to him through the sessions.

Hamid was a male student who was 8 years and 10 months old. He was diagnosed with mild autism and his PIQ was 80. Hamid was enrolled RLT School when he was 4 years of age. His stereotypic responses included hand flapping and head movement. Hamid has reading and writing skills, and his skill of understanding what he reads is improving. In terms of his expressive language, he is able to hold short conversations and usually used four-word sentences. His tablet skill was fine.

Layla, the only girl in the study, was 7 years and 5 months old at the start of the study. Her PIQ score was 72 with a moderate severe ASD. She was enrolled RLT School 4 years ago with a minimal language and a history of tantrums. Layla had stubborn behavior and sometimes exhibited crying and noncompliance. She also appeared to communicate with people she knew before.

Khalil was 6 years and 9 months old boy diagnosed with moderate autism. His PIQ score was 76. He has been a student in RLT for 2 years and so far he has difficulty with receptive and expressive language. He responded to familiar instructions and primarily used three-word sentences to communicate his needs. Khalil was good with tablets, but he presented poor fine motor skills while touching the screen.

2.3.2 Ethical approval

The study was approved by The Ethics board of Istanbul Technical University (Ref no. BM-28, dated May 12, 2015) (see Figure B.1). Written informed consent was signed for each child by their parents after they received a complete explanation on the purpose of the study and assurance about the confidentiality of the information (see Figure B.2). The form was also signed by the child's educator as a witness. Written permission was also obtained from the two schools where the experiments were conducted (Figure B.3 and Figure B.4).

2.3.3 Setting

The sessions in the experiments were conducted in the participants' respective school. The room was furnished with suitable chairs and tables. It was quiet and free from distractions to keep the participant's attention only on the application. As teaching

using software presented on computers or tablets were shown to increase children's attention and desire to learn in comparison to traditional programs (Lorah et al, 2013; McNaughton and Light, 2013), our application was presented on a tablet PC “Samsung Galaxy Note 10.1”. A digital camera was used to record some sessions for each participant and also some pictures were taken as in Figure. 2.1. Later, videotaped files were used to keep track of the child's behavior.

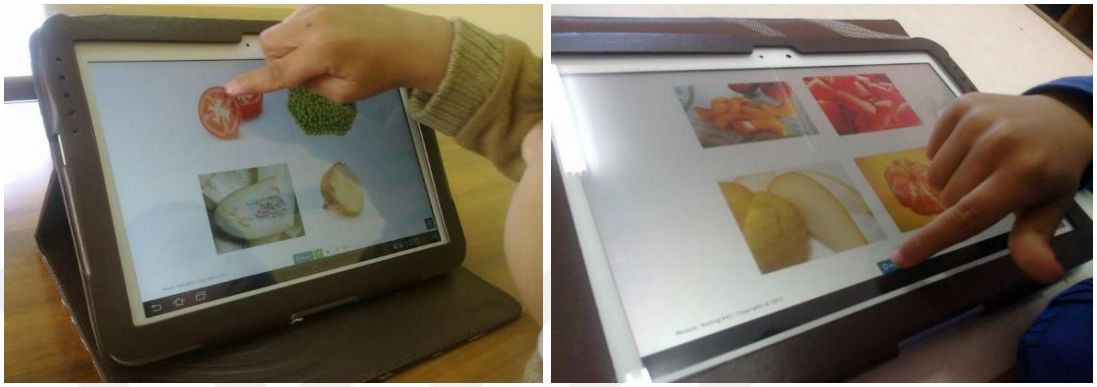

Figure 2.1 : Participant’s activities during the conduction of the experiments.

2.3.4 Target objects

We chose five common categories for a real-world dataset: fruits, vegetables, animals, tools for eating & cleaning and furniture. For each category, we picked six objects (see Table A.1). These objects were chosen according to two principles: 1) they represented the most common categories of concrete objects in the daily life of a child. 2) they were commonly used or were familiar to children with ASD. Children’s own teaching staff were consulted when the decision for objects and categories was made.

Picture stimuli of target objects were colored images, and they were collected using image search engines, in particular Google and Bing. Every object class contained about 31 different images and dataset has a total of 930 images. Some preprocessing was performed on the images to bring them to uniform resolution and size. We used an attribute-based approach (Farhadi et al, 2009) to select levels of difficulty: color, shape, size, plurality, cross-section, pattern, material (made of) and function (see Figure 2.2). Indeed, levels of difficulty did not have strict definitions and they depend on the nature of a category. In order to measure the ability of children to generalize (i.e. they do not memorize a particular picture, but learn the concept), different images were used for each phase. So we divided the dataset of images into seven groups to ensure that the images used in the assessment sessions of one phase will not be used

in the following assessment sessions for other phases. Moreover, the images used for teaching and assessment sessions of the same phase should be different, otherwise over-fitting could occur (Alpaydin, 2014). Some preprocessing were performed on the objects' images. We performed the following tasks: 1) Uniform the resolution of images and they scaled to be 300x360 pixels. 2) Minimize the size of the image files to be less than 50KB. 3) Crop unused and irrelevant parts of images or focus on the targeted objects.

Auditory stimuli consisted of recorded voice saying the name of the object such as 'apple', 'cat', 'table' or asking about the function of the object in the three categories (animals, tools and furniture) whose functions could not be expressed visually, using questions such as ''What do you sit on?'' or ''What do you use to eat soup?'' All auditory stimuli were of the same volume and intensity. The audio recordings were recorded by a native Arabic speaker female university student in a sound-proofed room and they were provided in mp3 format.

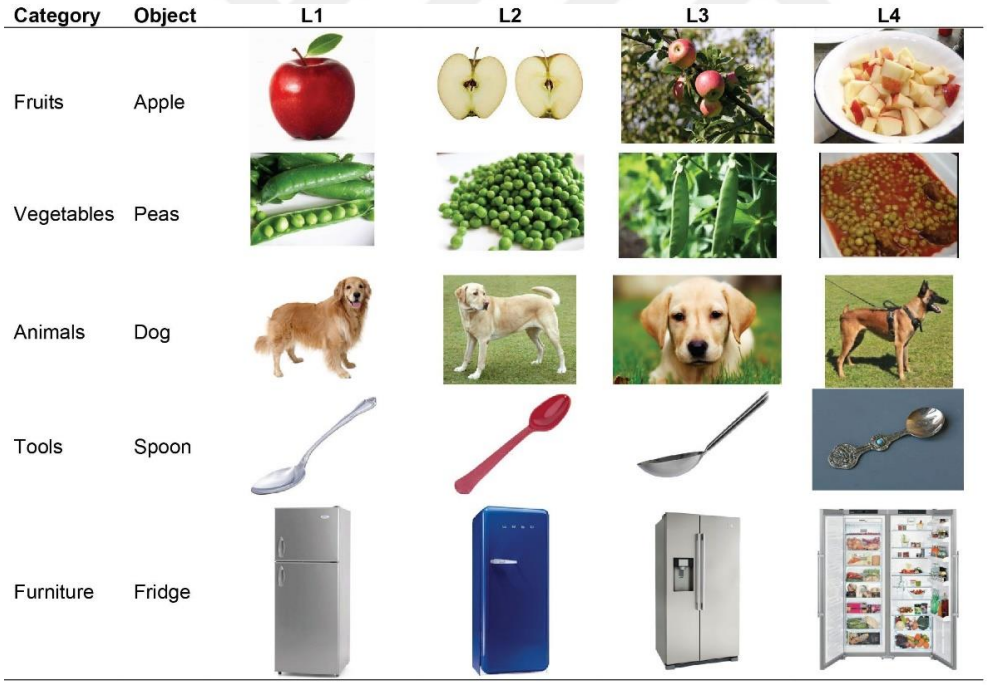

Figure 2.2 : A sample of images from the dataset used for teaching and assessment. Columns L1-L4 show different difficulty levels.

2.3.5 The web application

The method of identifying an object from an array after hearing its name is referred to as receptive labeling (Grow et al, 2014) or object label trial (Volkert et al, 2008). The application was constructed based on this method.

The procedure of an assessment trial was as follows: the application firstly presented an array of four objects (e.g., images of an apple, banana, pear and orange) on the tablet's screen as shown in Figure 2.3(a). Second, the participant immediately heard an auditory stimulus through the tablet. Third, the child engaged in a response (e.g., pointing to or touching the picture on the screen of the tablet). A child was encouraged to respond as quickly as possible and did not receive any feedback on the accuracy of his response.

In a teaching label trial, the application started in a way, which is similar to that in the assessment trial. If the child chose a correct answer, the application gave visual (smileys) and acoustical reinforcement (applause) to reward learning as shown in Figure 2.3(b). The application also provided the written name, picture and auditory stimulus for the correct answer. If the given answer was incorrect (i.e. one of the distractors is chosen), the application gave a certain acoustical sound to indicate the wrong answer and the child was asked to try again. An incremental cue was presented by shaking the correct answer horizontally two times in order to help the child. If the child's response was again incorrect, he was then given another opportunity to answer by shrinking three distractors in order to emphasize the correct answer. Finally, if the child could not answer correctly, the correct answer was given and shown on the tablet's screen.

Figure 2.3 : (a) Screen display of baseline, teaching and assessment trial. (b) The reward screen for the correct answer in a teaching trial. (c) Screen display of an initial assessment trial.

2.3.6 System architecture

The system was designed and built as a web-based application. That design decision allowed for more flexibility to access the application from anywhere and using any

device (laptop, tablet, or mobile). As a web application, it was built using a three tier architecture: Client/User, Server and Data tiers (see Figure 2.4). Many of the latest technologies were used in the development of our application. The front-end of the application was coded using HTML, CSS and JavaScript. HTML5 and CSS3 provided features such as responsive design and animations, which were used to create a child-friendly and interactive user interface for the application. Furthermore, the investigation was implemented as mini-SPAs (Single Page Applications) within the whole application using AngularJS (Url-1, 2015). As a JavaScript framework, AngularJS provided more flexibility and reusable components throughout different parts of the User Interface. The back-end of the application relied mainly on Microsoft SQL Server for data storage and retrieval in addition to ASP.NET MVC 5 for data manipulation and communication with the client. Hence, providing a separation of concerns between different layers of the application.

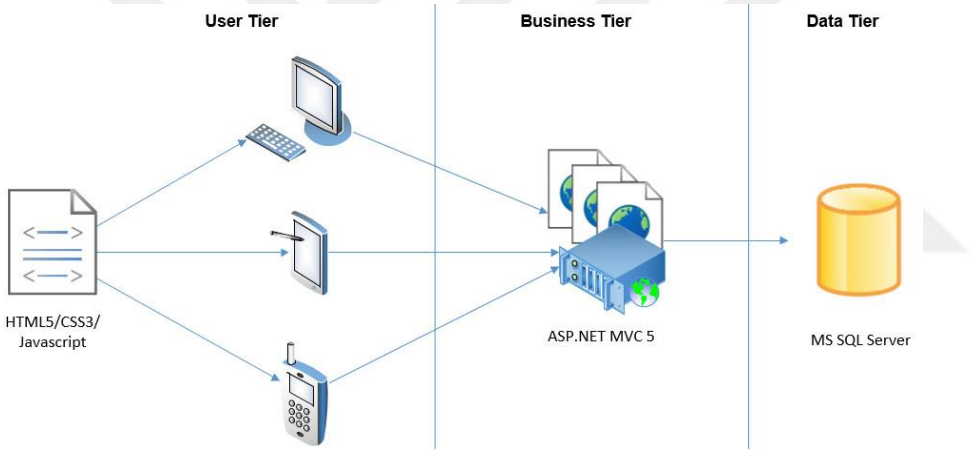

Figure 2.4 : System architecture of the web application.

2.3.7 Experimental design

This study employed an alternating treatments single subject design (Barlow and Hayes 1979; Horner et al., 2005) with an initial baseline phase and a final "best treatment" phase (Tincani, 2004; Devlin et al, 2011). Baseline data were collected to strengthen the study by showing the rates before treatment sessions began. Following baseline, PL and AML interventions were presented and alternated across sessions (Lorah et al, 2013). The data from the alternating treatments phase for each participant demonstrated which treatment was more effective. The most effective treatment was applied alone as the final phase for a number of assessment sessions. We chose the alternating treatments design since it meets two objectives. First, it ensured that a

change in child's performance is a result of the interventions rather than another event occurring at the same time. Second, an alternating treatments design is efficient for comparing effects of two or more treatments.

2.3.8 Dependent measures

The dependent variables were the percentage of correct responses (accuracy) and mean response latency (RL). Accuracy and mean RL for all trials were computed for each participant during the sessions. The number of auditory repeats (AR) was also recorded. All of these measurements were recorded automatically by the web application for each trial during teaching and assessment sessions. Measurement results for each session and trial were reported by the application.

2.3.9 Procedures

There were four phases in the current investigation: initial assessment, baseline, intervention and final phase. The design of the investigation is summarized in a flowchart in Figure 2.5.

Initial Assessment. It was conducted to identify a prior probability of objects and categories for each participant. Five images for each object (total 150 images) were shown randomly with auditory stimuli to each participant. It required only yes/no type binary feedback (see Figure 2.3(c)). There was no help when the child was not able to select the correct answer. The probability of an object was computed as the proportion of successful trials:

$$P(O_i) = \text{No. of successful trials / total number of trials.}$$

Object's probability list was created for each participant based on his or her responses. Figure C.1 shows a sample of the initial probability list for participant 1. The probability for each category was also computed as the average of its objects' probabilities.

Baseline. It consisted of ten assessment sessions, each comprised 30 label trials using only one level of difficulty. Figure 2.6 shows the flowchart of the assessment session. The trials were presented sequentially without replacement. Baseline phase served to evaluate child's performance on each level prior to the comparison and it also provided data in order to be able to apply the next phase. Categories and objects were presented to each participant in baseline sessions based on their probabilities so that category

and object with high probability was shown first. If there was no probability list, we used the arrangement that we prepared in advance within the application preferences. When a student completed the assessment session successfully, the number of correct and incorrect responses, response latency and number of auditory repeats were automatically produced. When children played the assessment sessions for level L_1 and L_2 first time, we noted that some of them recognized objects by their color. Because of that, we added two extra assessment sessions. We focused on the shape of the object, and the goal was to determine if the child could easily recognize objects from the outline drawings of their shapes. Figure 2.7 shows a sample of outline images.

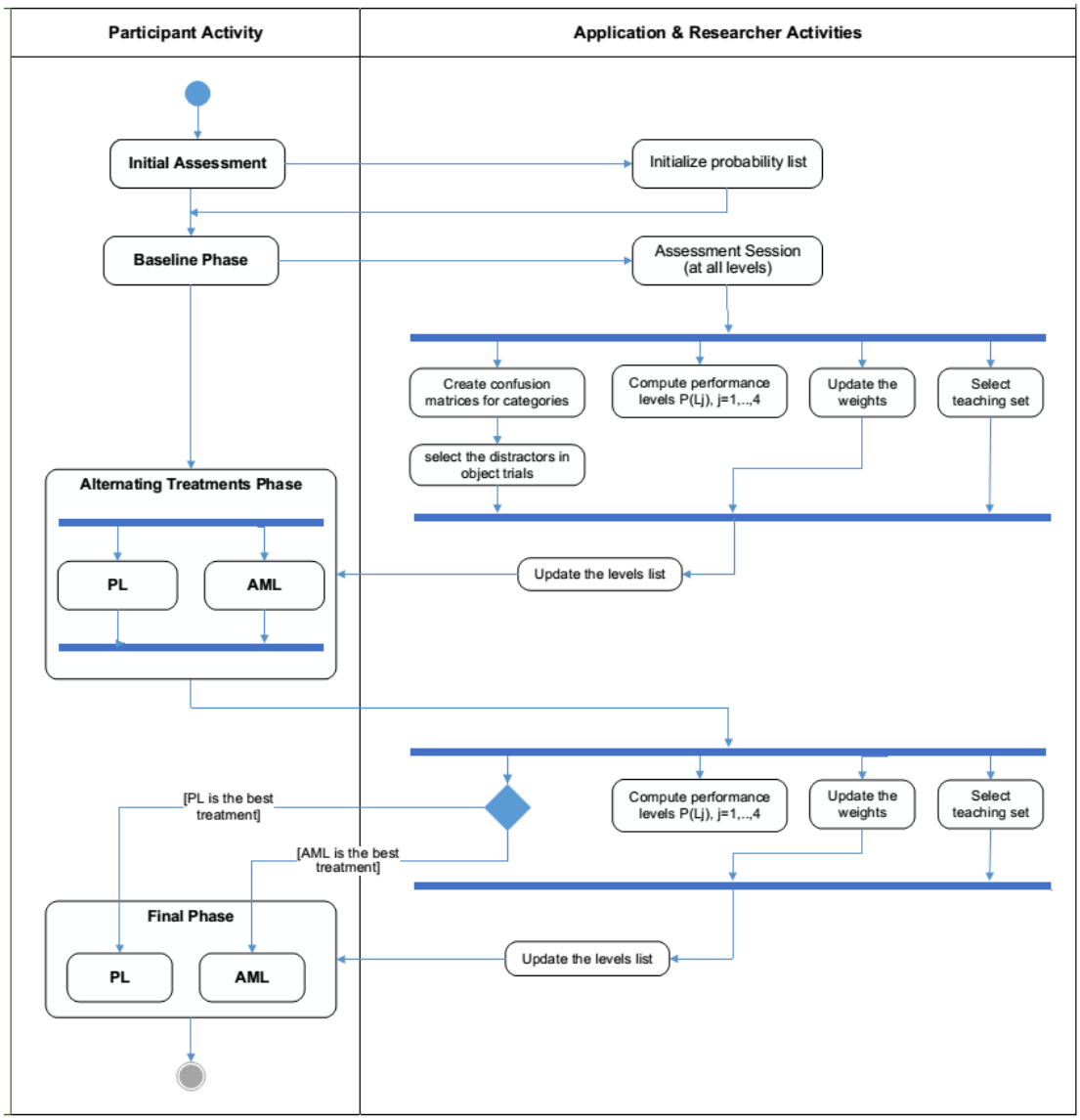

Figure 2.5 : Flowchart of the investigation for a participant.

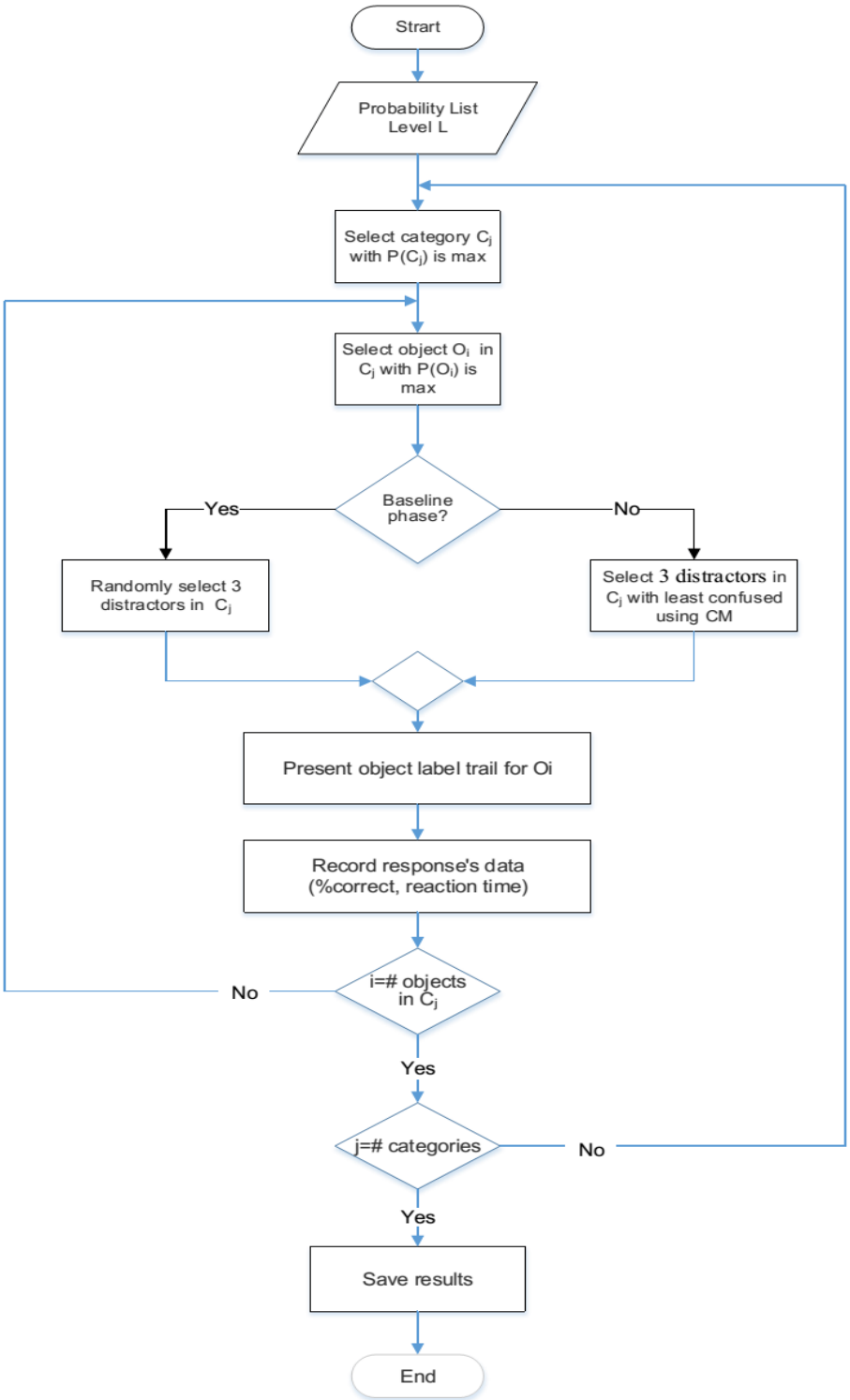

Figure 2.6 : Flowchart of the assessment session in baseline phase.

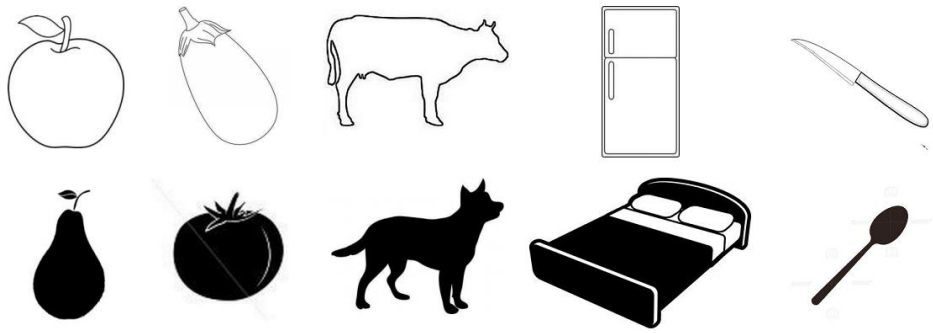

Figure 2.7 : Sample outline of drawing images.

At the end of each iteration, the following steps were performed for each participant:

i) Computed the performance at each level L_j as follows:

$$P(L_j) = \frac{1}{N}\sum_{i=1}^{N} RC(i,j),\ j = 1,..,4$$

ii) Updated the weights according to equation (2.2).

iii) Selected queries or teaching set for each level using our approach WRAML.

Intervention. Alternating treatments phase (ATP) was implemented across eight sessions. Treatments were counterbalanced across sessions. We conducted the treatments two to three times per week. In this phase, difficulty levels were adjusted based on the participant's baseline performance using the method described above in formulation. ATP consisted of two treatments:

PL treatment which consisted of one teaching condition followed by one assessment condition in the same session as shown in Figure 2.8.

- Teaching condition: five teaching label trials from each included levels where the objects of the trials were randomly selected (one from each category).
- Assessment condition: ten assessment trials from each included levels where the objects of that trials were selected without replacement (two for each category).

AML treatment which consisted of four blocks (if four levels were included), with each block divided into a set of n teaching trials followed by 10 assessment trials (see Figure 2.8).

- Teaching condition: a batch of n teaching trials from one level where trial's object was selected based on our query selection approach of the previous phase (n= 4, 3 or 2).
- Assessment condition: same as in the PL treatment.

When the AML session is started, the screen in Figure 2.9(a) is shown. We should select the levels that included in the session and groups of images for training and test. Then we should select a batch of n queries from each level included in this session (Figure 2.9(b)).

PL session	
Teaching condition: 5 trials from each included level	Assessment condition: 10 trials from each included level

AML session							
Teaching n trials L1	Assessment 10 trials L1	Teaching n trials L2	Assessment 10 trials L2	Teaching n trials L3	Assessment 10 trials L3	Teaching n trials L4	Assessment 10 trials L4

Figure 2.8 : Session schedule of intervention for PL and AML treatments.

Final phase. The most effective treatment "best treatment" was applied alone as the final phase for two or three sessions. The best treatment (PL or AML) and difficulty levels included with the interventions were selected based on the participant's performance in ATP phase. The final phase served to determine if performance was influenced by the best treatment or whether other procedures could affect child's performance.

(a) (b)

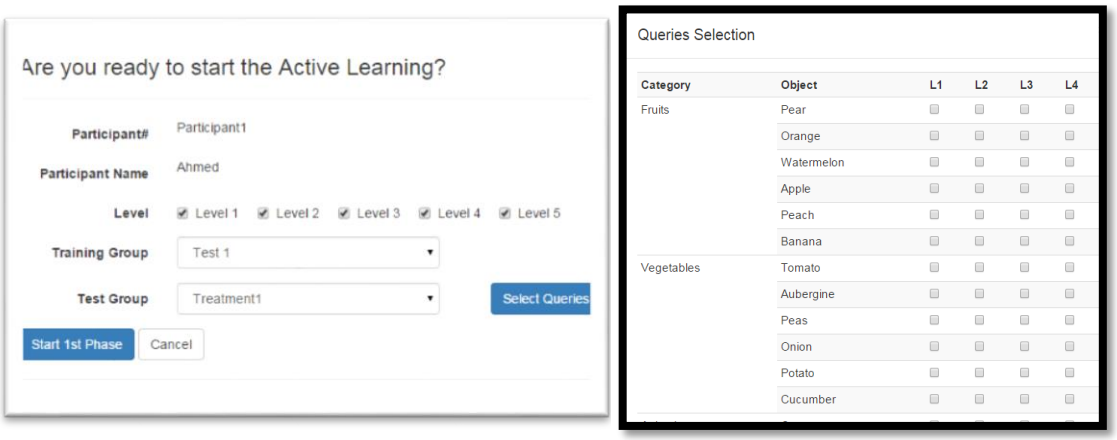

Figure 2.9 : Screens display of the AML session.

2.3.10 Inter-observer agreement and procedural Integrity

The data were collected by the web application and all measurements were recorded automatically by the application for all trials of the sessions. To ensure the reliability of the collected data, an independent observer who was the child's educator selected some videotapes of the participant's sessions and checked the measurements. Inter-

observer agreement between the experimenter and the independent observer was 100% across 19% of the sessions. To assess procedural integrity, the independent observer checked whether or not the experimenter had implemented all phases of the investigation in its correct sequence with the same number of sessions for each participant. This was easily done by reviewing the results page in web application. Procedural integrity was reported 95% agreement across all sessions (number of sessions for final phase was not equal for all participants).

2.4 Results

The investigation was conducted using web application that was presented on a tablet PC. The data were also collected by the same software. Figure 2.10 shows a sample of the detailed results' page of the assessment session. We collected data over an eight-week period in two schools. Data obtained from the study were analyzed by visual and statistical analysis methods. The visual analysis included examining data for accuracy and mean RL. Line graphs were used to show participant performance levels. The degree of the effect produced by interventions is determined by the amount of vertical difference between data paths (Cooper et al, 2007). To decide which PL or AML was more effective, the treatment with high accuracy and low mean RL during ATP being the best. However, when the treatment with faster response has lower accuracy or vice versa, it is difficult to reach a convincing conclusion. The inverse efficiency score (IES) is a way to combine both measures (accuracy and RL) (Bruyer and Brysbaert, 2011) and it is commonly used in case of speed–accuracy trade-off. The IES is calculated by dividing correct response latency by accuracy within condition; a lower score means better performance.

P#	Name	Group/Level	#Questions	Accuracy	%Correct	Finish Date	Finished?	
2	Omar	Test 2/5	30/30	21/9	70%	25/04/2015	Yes	Visual Results

Q#	Correct?	Answer	Response	Choices				Response Latency	Repeat Count
1	Yes	Orange5	Orange5	**P1:**Pear5	**P2:**Orange5	**P3:**Watermelon5	**P4:**Apple5	22	1
2	Yes	Banana5	Banana5	**P1:**Orange5	**P2:**Peach5	**P3:**Banana5	**P4:**Watermelon5	4	0
3	No	Peach5	Apple5	**P1:**Apple5	**P2:**Pear5	**P3:**Peach5	**P4:**Banana5	4	0
4	No	Pear5	Watermelon5	**P1:**Watermelon5	**P2:**Pear5	**P3:**Orange5	**P4:**Banana5	3	0

Figure 2.10 : A sample of details result of an assessment session.

2.4.1 Ali

Figure 2.11 and Figure 2.12 represent the percentage of correct responses and mean response latency during baseline, alternating treatments and final best treatment phases for Ali. During the baseline sessions at levels ($L_1 - L_3$), Ali demonstrated high accuracies (range, 90–100%), stable and low mean RL (range, 3.37–3.9s). However, the decreases were observed in performance for sessions at level L_4 where the average of correct responses was 76.5 and average mean RL was 6.58 (range, 6.33-6.83s). Average performance for all levels across baseline and ATP phases are shown in Table 2.3. Ali has passed L_1, so it was not included in the following phases.

During the ATP phase, the PL and AML were alternated across eight sessions. The data paths show a higher accuracy and lower RL when the AML was in place IES score in Table 2.4 demonstrated that AML was better than PL. As shown in Table 2.3, Ali met the criterion for passing additional two levels (L_2 and L_3) during this phase. AML was the best treatment and only AML was implemented in the final phase. The data of the best treatment phase were collected during three sessions for AML and averaged 97% (accuracy) whereas the increases were observed in mean response latencies because only the most difficult level L_4 was included in this phase.

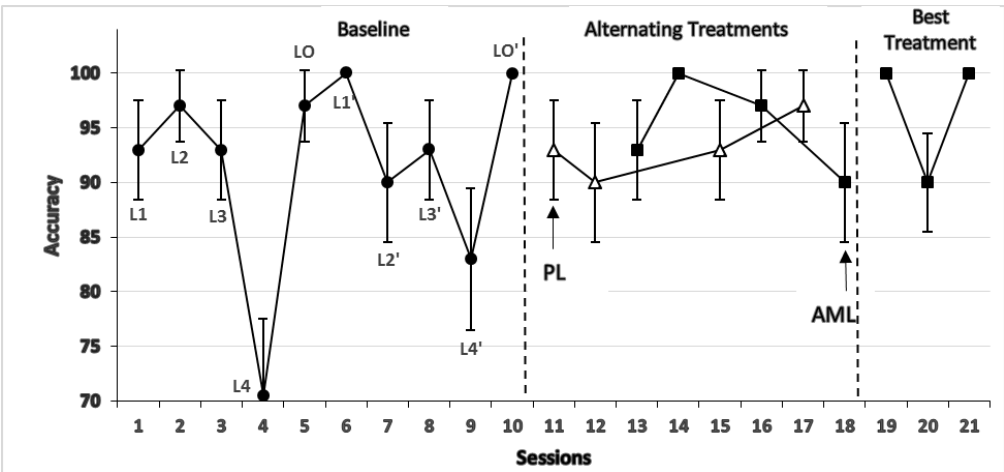

Figure 2.11 : Accuracy during baseline, ATP and final phases for Ali.

Table 2.3 : Mean performance for all levels across baseline and ATP phases (standard error of the means are given in parentheses).

Participant		Baseline Phase				ATP Phase			
		L_1	L_2	L_3	L_4	L_1	L_2	L_3	L_4
Ali	Accuracy (%)	96.7(2.3)	93.3(1.8)	93.3(3.2)	76.7(5.5)	Pass	99 (1.3)	96(2.1)	88(3.7)
	Mean RL (s)	4.12	4.42	4.16	6.96		4.32	4.21	5.68
Osman	Accuracy (%)	80.0(5.2)	70.0(4.2)	73.3(5.7)	76.7(5.5)	84(4.2)	79(4.6)	75(4.9)	71(5.1)
	Mean RL (s)	6.80	8.87	9.39	10.40	5.77	7.00	7.54	9.58
Hamid	Accuracy (%)	98.3(1.7)	93.3(2.0)	96.7(2.3)	86.7(4.4)	Pass	98(1.8)	Pass	88(3.7)
	Mean RL (s)	3.51	3.39	3.54	7.82		3.38		6.15
Layla	Accuracy (%)	78.3(5.4)	75(4.0)	81.7(5.0)	68.3(6.1)	90(3.4)	94(2.7)	89(3.6)	75(4.9)
	Mean RL (s)	3.93	3.65	4.05	6.36	4.12	4.43	4.16	5.76
Khalil	Accuracy (%)	91.7(3.6)	88.3(2.9)	85(4.6)	73.3(5.8)	94(2.8)	95(2.8)	80(4.7)	70(5.5)
	Mean RL (s)	6.89	8.35	7.17	13.66	8.61	10.34	11.70	17.18

Note: highlighted cell indicates that the participant has passed this level.

Table 2.4 : Average accuracy, mean of correct response latency (CRL) and the inverse efficiency score (IES) for treatments PL and AML during ATP phase.

	PL			AML		
Participant	Accuracy (%)	CRL	IES	Accuracy (%)	CRL	IES
Ali	93.3 (2.3)	4856	52.0	95.0 (2.0)	4156	43.7
Osman	73.8 (3.5)	7456	101.1	80.6 (3.1)	6833	84.7
Hamid	90.0 (3.4)	4531	50.3	95.0 (2.5)	4233	44.6
Layla	83.9 (2.9)	4321	51.5	89.9 (2.4)	4553	50.7
Khalil	85.6 (2.8)	10110	118.1	84.2 (3.3)	10276	122
Overall	85.3 (3.0)	6254.8	74.6	88.9 (2.6)	6010.2	69.2

standard error of the mean are given in parentheses.

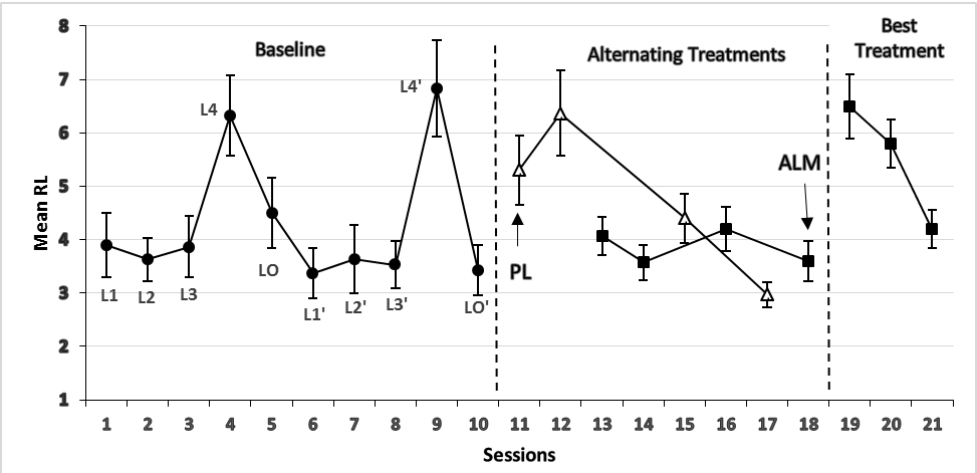

Figure 2.12 : Mean RL during baseline, ATP and final phases for Ali.

2.4.2 Osman

Figure 2.13 and Figure 2.14 represent the percentage of correct responses and mean response latency during baseline, alternating treatments and final best treatment phases for Osman. During the baseline sessions, Osman demonstrated moderate accuracies, he averaged 75% (range, 67–80 %), meanwhile he demonstrated high mean RL, his average RLs was 8.53s (range, 6.37–13s). As depicted in Table 2.3, Osman could not pass any levels during baseline.

During the ATP, the PL and AML were alternated across eight sessions. Figure 2.13 demonstrated stable data point and a little overlap in data paths. A higher accuracy and lower RL are shown with AML treatment. As AML was significantly lower IES score than PL (see Table 2.4), AML was the best treatment for Osman. For the final best treatment phase, the data were collected during two sessions for AML and averaged 83.5% with no changes in average for mean RL.

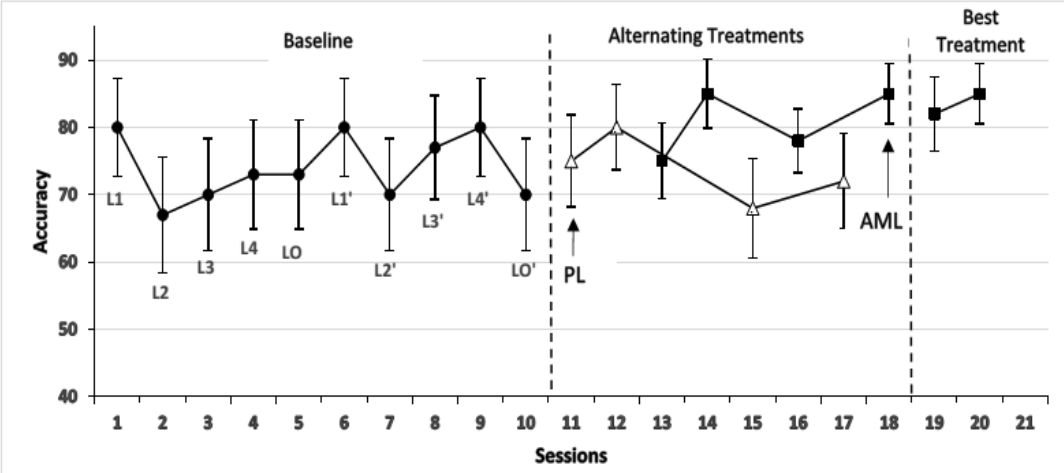

Figure 2.13 : Accuracy during baseline, ATP and final phases for Osman.

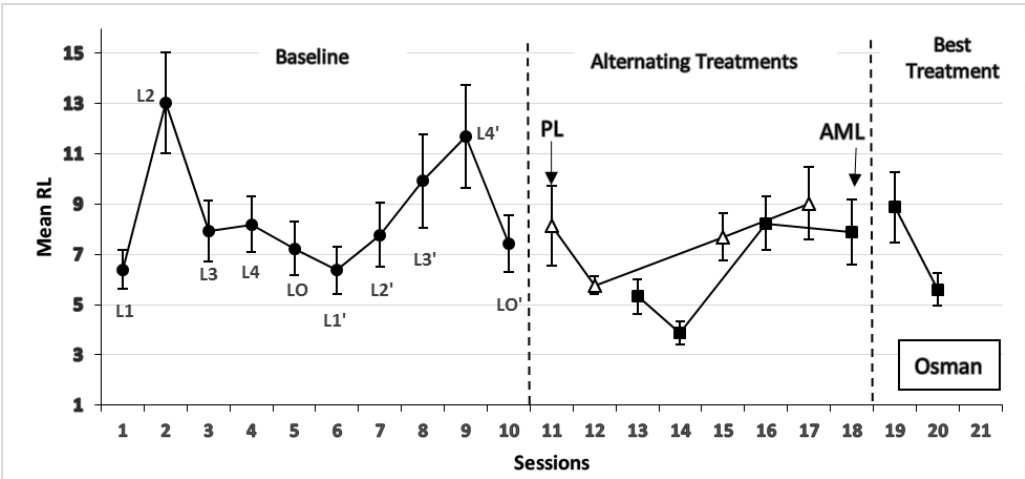

Figure 2.14 : Mean RL during baseline, ATP and final phases for Osman.

2.4.3 Hamid

Figure 2.15 and Figure 2.16 represent the percentage of correct responses and mean response latency during baseline, alternating treatments and final best treatment phases for Hamid. During the baseline sessions at levels ($L_1 - L_3$), Hamid demonstrated high correct responses (range, 93–100 %), stable and low mean RL (range, 2.6–3.67s). However, decreases were observed in performance for sessions at level L_4 where average correct responses was 87% and average mean RLs was 7.4 (range 7.03-7.77s). As depicted in Table 2.3, Hamid has passed two levels during baseline, so those levels were not included in later sessions.

During the ATP, the PL and AML were alternated across eight sessions. The data paths show a lower RL when the AML was in place. Based on IES score in Table 2.4, AML

was the best treatment for Hamid and it was implemented in the final phase. Moreover, Hamid has additionally passed level L_2 during this phase. The data of best treatment phase were collected during three sessions for AML and average accuracy was 97% with no changes in average for RL.

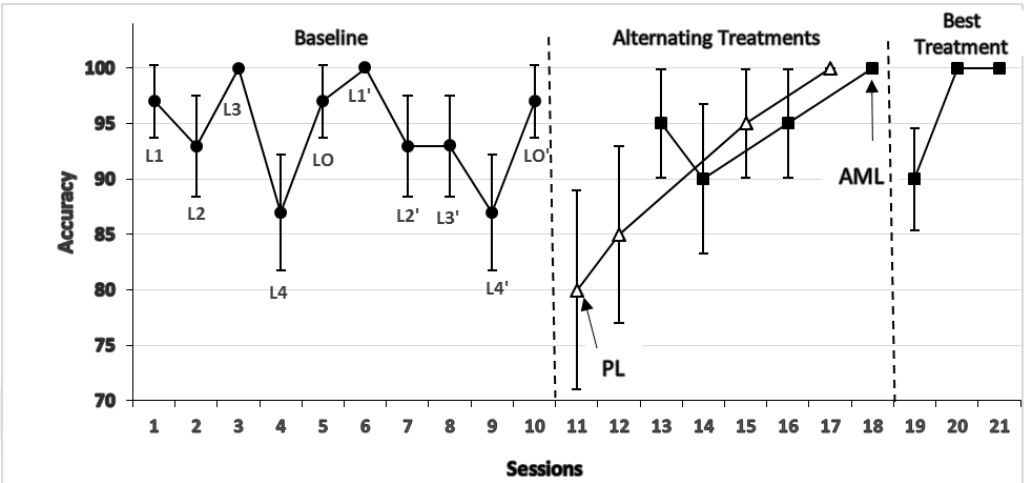

Figure 2.15 : Accuracy during baseline, ATP and final phases for Hamid.

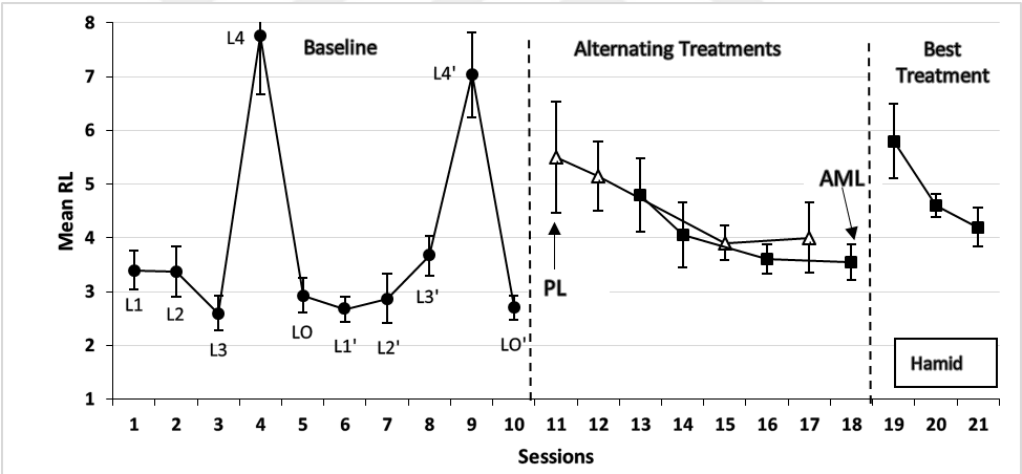

Figure 2.16 : Mean RL during baseline, ATP and final phases for Hamid.

2.4.4 Layla

Figure 2.17 and Figure 2.18 represent the percentage of correct responses and mean response latency during baseline, alternating treatments and final best treatment phases for Layla. During the baseline sessions, Layla demonstrated moderate correct responses, she averaged 75.9% (range, 67–90 %), meanwhile she demonstrated relatively low mean RL, her average RLs was 4.07 (range, 2.83–6.23s). Layla could not pass any levels during baseline phase.

During the ATP, the PL and AML were alternated across eight sessions. The data paths show clearly a higher accuracy and higher mean RL when the AML was in place. As shown in Table 2.3, Layla met the criterion for passing L_2 during this phase. Although AML was slightly lower IES score than PL (see Table 2.4), AML only was implemented in the final phase and three levels were included. During this phase, the data were collected during three sessions for AML and averaged 92.3%, whereas the increases were observed in mean RL which average was 5.9s.

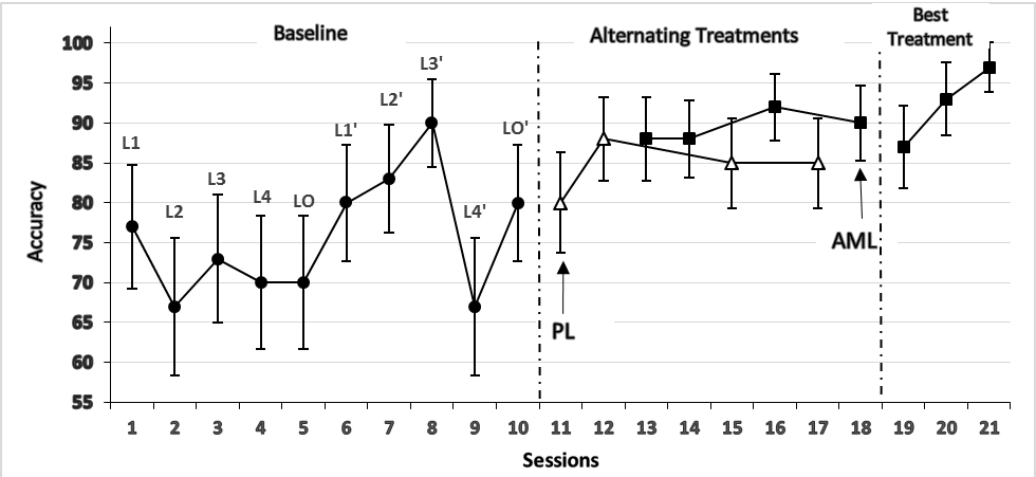

Figure 2.17 : Accuracy during baseline, ATP and final phases for Layla.

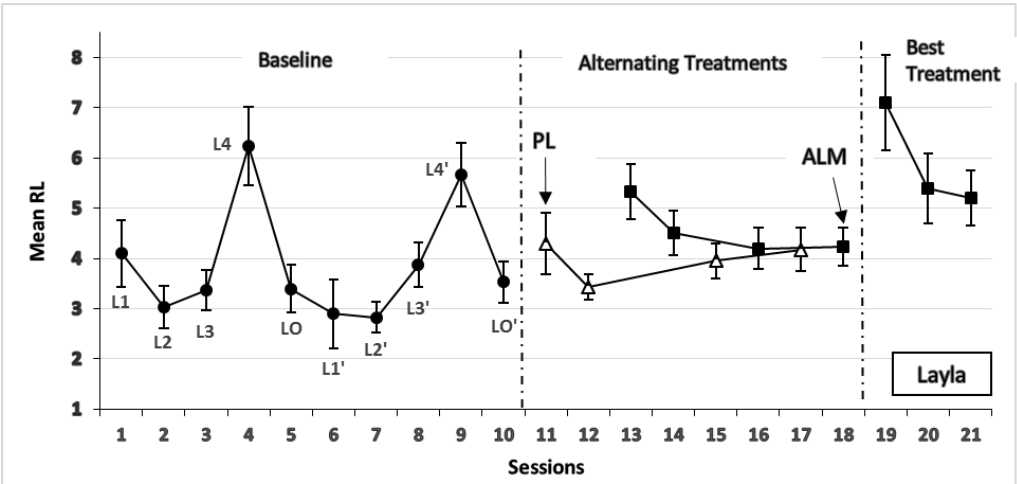

Figure 2.18 : Mean RL during baseline, ATP and final phases for Layla.

2.4.5 Khalil

Figure 2.19 and Figure 2.20 represent the percentage of correct responses and mean response latency during baseline, alternating treatments and final best treatment phases

for Khalil. During the baseline sessions at levels ($L_1 - L_3$), Khalil demonstrated relatively high correct responses (range, 83–97 %), high mean response latency (range, 5.37-9.3s). However, the decreases were observed in performance for sessions at L_4 where average of correct responses was 73 and the average mean RLs was 13.27. No level was passed in this phase.

During the ATP, the PL and AML were alternated randomly across seven sessions (absent in session 18 for AML). Since there was an overlap between data paths, we could not determine the effect produced by two interventions. Based on IES score in Table 2.4, PL was the best treatment and it was implemented in the final phase. As shown in Table 2.3, Khalil met the criterion for passing two levels (L_1 and L_2) during this phase.

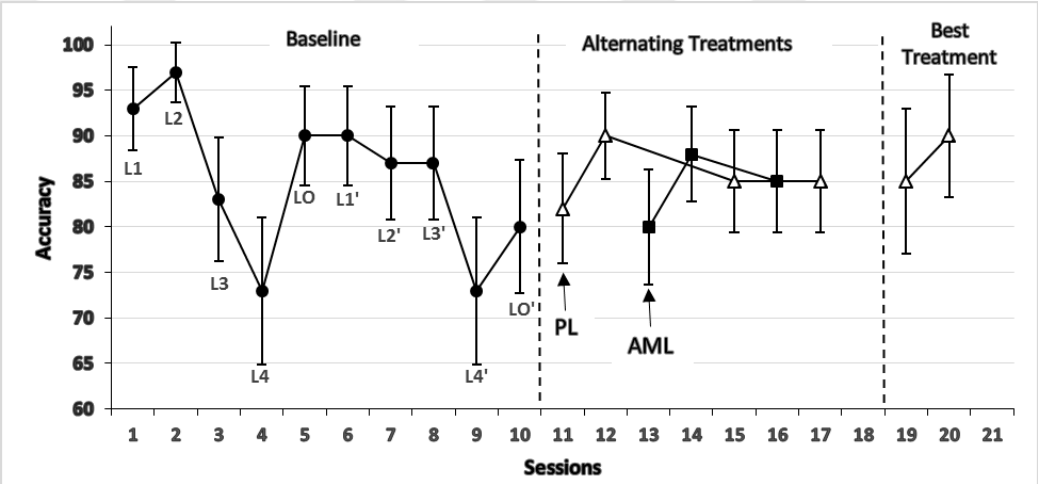

Figure 2.19 : Accuracy during baseline, ATP and final phases for Khalil.

The data of best treatment phase were collected during two sessions for PL and averaged 87.5% whereas the increases were observed in mean RL which average was 12.55s, because the most difficult levels L_3 and L_4 were included in final phase.

2.4.6 Reduction in required teaching trials

In the ATP phase, the number of teaching trials for a participant was fixed for the PL method (five for each included level) wherever for the AML method it was set from four to three in ATP and two in the final phase. During the ATP, as shown in the figures above and Table 2.4, AML had higher accuracy than PL for all participants except Khalil. Moreover, participants 1, 2 and 3 responded more quickly with AML. Further analysis showed that to achieve an average accuracy of 85% across all

participants during ATP phase, PL needed 20 teaching trials, while AML needed only 12 to achieve an average accuracy about 89% (see Table 2.4). This means a 40% (12/20) reduction in teaching effort. In the final phase, AML was implemented for four participants and achieved average accuracy of 90%. Thus, we can conclude that AML required fewer teaching trials than PL in order to achieve a similar or higher performance. Reduction of the number of instances required for training is one of the advantages of AML.

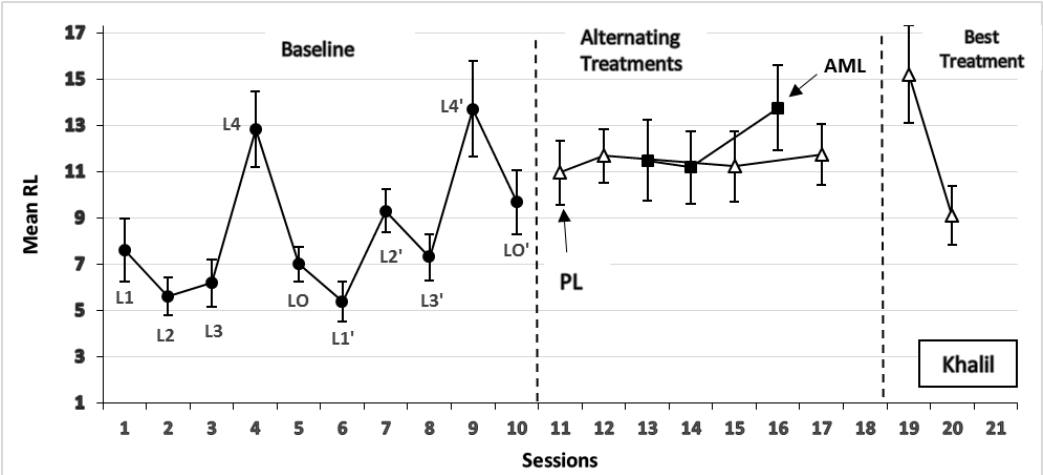

Figure 2.20 : Mean RL during baseline, ATP and final phases for Khalil.

2.5 Discussion

Our work represents the first attempt to use ML techniques for teaching children with ASD. To the best of our knowledge, there is no similar framework so far. This study designed to investigate the use of AML for teaching children with ASD, and compared the effectiveness of PL and AML on teaching object recognition for those children. Experimental results on participants showed that compared to PL, AML approach was generally more effective in terms of accuracy. One-way ANOVA (with α=0.05) was used to test the significance of differences between the treatments' means. The results, $F(1) = 5.143; \mathrm{p} = 0.023$, demonstrated that there was a statistically significant difference in accuracy level between the means of PL and AML. Our approach and procedures provide two features that helped to reduce repetition in learning environment: 1) Minimizing the number of teaching trials required for training. 2) Determining mastery criterion for levels. When a participant reached mastery criterion, the application no longer assesses this level in the following phases. This result is

consistent with a recent study (Harris et al, 2015) which found that reducing repetition in teaching may enhance learning in individuals with ASD.

2.5.1 The suitability of AML framework for students with ASD

The proposed AML described in this thesis can be applied to other education domains. However, teaching children with ASD is a particularly suitable domain for our framework for two reasons. First, teaching students with ASD in the classroom or at home is a challenging due to the diverse needs and sensory issues faced by the student. They highly depend on their parents and teachers. Thus, the reduction in teaching effort (with less number of trials) is an important consideration for the time being. Second, the AML method aims to design an optimal teaching session for individual children. This approach will improve the teaching process and personalize learning for each child.

2.5.2 Efficiency of the teaching condition

The investigation also measured the efficiency of the teaching condition on participant's performance. The average accuracy at only included levels in all three phases (baseline, ATP and final) are shown in Figure 2.21. The results show that each participant improved from baseline phase to the final phase. For example, Ali's accuracies increased from an average of 76.7 to 96.7. Moreover, the standard error of the mean improved for the first four participants. Since the mastered levels are eliminated and only the most difficult ones are included, the application complexity will increase when the participant goes through from baseline to ATP and then to final phase. In spite of this complexity increasing, all participants acquired their highest average accuracies in the ATP and final phases. These increases can be attributed to the effectiveness of the teaching condition which was designed according to the ABA procedures. This finding is consistent with previous researches showing that ABA-based treatments for students with ASD have been successfully used to help many individuals with ASD (Walsh, 2011; NAC, 2015).

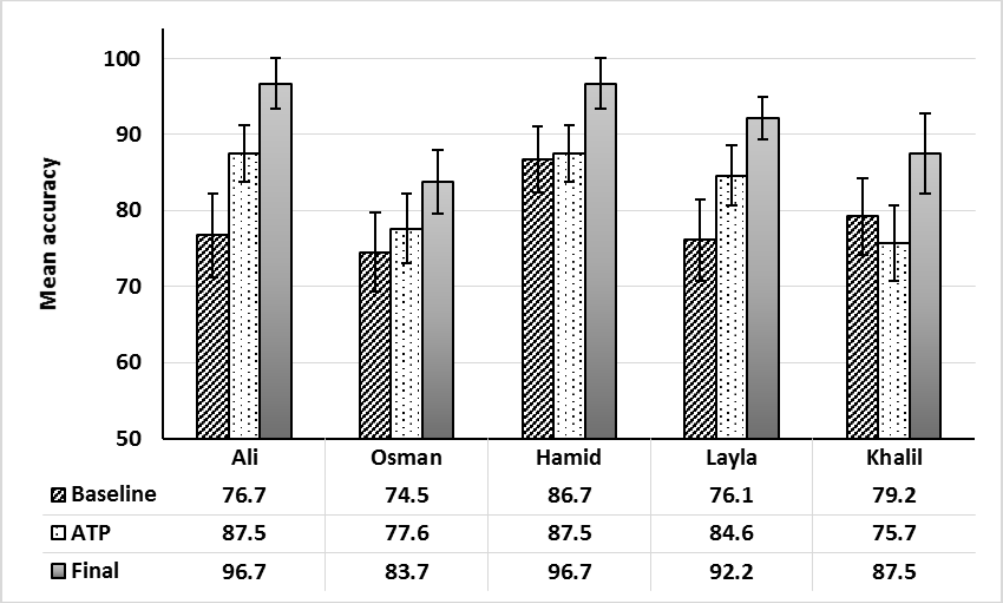

Figure 2.21 : Mean accuracy at only included levels in all phases. Error bars are standard error of the means.

2.5.3 Limitations and future directions

There are several limitations of this study that should be considered in relation to the findings. First, the small number of participants limits the generalization of the findings to other children with ASD. In addition, a small number of objects and categories used in the study leads to a similar limitation, because small number of teaching and assessment instances reduces the reliability of the results. Moreover, the objects in our investigation are commonly used and familiar to children with ASD. Future research should include a larger sample of individuals with ASD with a wide variety of ability levels using more and unfamiliar stimuli. Second, we got our findings when PL was applied first in the ATP. The question is whether we will get the same results in case the AML was applied first. This point needs to be examined in future studies. Third, even though the findings were replicated across four participants in terms of accuracy, these results should be interpreted cautiously in terms of the mean RL. For example, Layla had higher mean RLs with AML than with PL. We found that the mean RLs was affected by the level of the participant's tablet usage skills and level of engagement. In future studies, participants might either be required to have a set of prerequisite fine tablet skills or they may be provided with an initial independent set of sessions that aim to increase the tablet usage skills. We observed that Osman and Khalil had poor skills to touch screen and to select the correct picture, and therefore this delay affected their response latencies. We also observed that participants were

inactive while waiting to download of the pictures for the next trial. Fourth, an estimated 30% of individuals with ASD are minimally verbal and display high levels of symptom severity and impairment of motor skills (Tager-Flusberg and Kasari, 2013). They remain minimally verbal, even if they have received years of interventions and a range of educational opportunities (Tager-Flusberg and Kasari, 2013). Since their ability to understand speech is often poor, in some cases these individuals just look at the pictures on the tablet's screen and listen passively to the auditory stimuli. In other cases, they look longer when matching pictures for sounds that they know. However, they fail to do so for sounds they do not understand (Edelson et al, 2008). For the above reasons, our findings cannot be generalized to children with severe ASD (typically minimally verbal). Fifth, selection of teaching set of objects for each level using our approach WRAML is mainly based on weighted error E_w which is affected by weight initialization and learning rate η. Determining the optimal initial weights of our framework definitely will enhance the learning performance. However, this task is hard because we have to reapply our experiment many times with all participants. Lastly, images of dataset in the current study were classified into four difficulty levels according to some attributes. This classification of difficulty levels was based on the researcher and educator’s experience. Future studies should consider other methods to choose the levels and evaluate the extent of difficulty.

3. OBJECT CATEGORIZATION TASK FOR STUDENTS WITH ASD

3.1 Using Assistive Technology to Enhance Teaching

Several studies show that computers and other technological devices are the preferred medium that provide a motivated learning process to most individuals with ASD (O'Malley et al, 2013; Smith and Sung, 2014). In comparison to traditional programs, teaching that uses software presented on computers or tablets were shown to increase attention and desire to learn for students with ASD.

In recent years, many empirical studies utilized a tablet or iPad as learning tools in educational program for students with ASD. Artoni et al. (2011) designed an instructional software for low-functioning autistic children based on the principles of ABA. The software aimed to create eLearning environments (didactic programs and monitoring learning) for teaching efficiently and effectively. The key feature of this software is its ability of collaboration with therapists, parents and caregivers to ensure that children's individual needs are completely addressed. Kagohara et al. (2012) aimed to teach two students with ASD to check the spelling of words using the spell-check function on common word processor programs. The results indicated that video-modeling intervention on an iPad improved teaching spelling skills to the children. Two children (aged 3 and 7) showed an increase in academic engagement and decrease in problem behaviors when teaching was delivered using an iPad compared to using a traditional approach in (Neely et al, 2013). Smith et al. (2013) investigated the effect of the present slideshows on a tablet to teach science terms to three adolescents with ASD. All students generalized new terms to activity sheets. Teachers and students agreed that the tablet was an effective and appropriate teaching tool. Doenyas et al. (2014) performed the first study on three Turkish children with ASD to observe their reactions to the tablet application and its effectiveness in teaching sequencing skills. In this study, all participants demonstrated improvement from initial to the final testing sessions. However, the application was not enough to teach the skills to youngest boy, who required external help, and it was so easy for the oldest boy, who did not use the rewards and he got bored early.

This study attempts to expand previous studies by using a tablet PC with a web-based application for teaching students with ASD. Due to several features for tablet PC, we suggest that the tablet is an effective educational tool to enhance teaching for the students with ASD. Moreover, the software allows us to analyze the data gathered during the learning sessions and make useful decisions about teaching process as we will see in the next section. The collected data analyzed using statistical packages R Commander (Rcmdr) (Fox and Carvalho, 2012).

3.2 Results and Discussion

3.2.1 General performance of students

Table 3.1 gives the comparison of participants' mean responses over all sessions. Participant 3 has the best performance in terms of accuracy and mean RT. IES is better to identify to the overall performance for other participants. According to Figure 3.1 with Table 3.1, when the accuracy is considered, participant 1 and 3 performed the best with respect to accuracy (92.9% and 93.9%, respectively) and mean RT (4.807 and 4.541, respectively). They have the highest accuracy and lowest mean RT while Participant 2 performed the worst. Participant 4 and 5 have the average performance. On the other hand, when the IES is considered, participants respective performance change for some participants.

Table 3.1 : The performance of participants. Mean response times (RT); mean of correct response times (PRT); mean of incorrect response times (NRT); inverse efficiency score (IES); percentage of auditory repeats (AR); number of trials (NT); student's overall performance. Standard error are given in parentheses.

Std. ID	Accuracy (%)	RT (sec)	PRT (sec)	NRT (sec)	IES (sec)	AR (%)	Overall Perfor.
1	0.929 (0.011)	4.807 (0.157)	4.474	8.574	4.812	0.118	best
2	0.766 (0.016)	8.162 (0.291)	7.566	10.110	9.881	0.236	worst
3	0.939 (0.011)	4.541 (0.157)	4.167	9.822	4.439	0.065	best
4	0.828 (0.014)	4.761 (0.133)	4.467	6.179	5.394	0.121	average
5	0.853 (0.014)	10.31 (0.353)	9.269	11.425	10.863	0.187	worst

3.2.2 Relationships between variables

Multiple linear regression analyses over 3090 trials of all five participants was conducted to test the correlations among three variables; response time as the dependent variable with accuracy and auditory repeats as the independent variables.

This analysis revealed a strong negative correlation between response time and accuracy, $\beta = -3.19$; $p < .0001$; whereas a significant positive correlation between response time and auditory repeats, $\beta = 8.95$; $p < .0001$. In other words, the higher participants' accuracy, the faster they were, and the higher participants' auditory repeats the slower they were. Figure 3.2 shows the model residuals against the fitted values. The line drawn on the plot is a linear least-squares fit the model. As predicted, there was a negative correlation between accuracy and auditory repeats, $\beta = -0.123$; $p < .0001$, i.e., the higher student's accuracy is, the fewer auditory repeats student needs. In sum, the regression analyses matched the the results we summarized in Table 3.1.

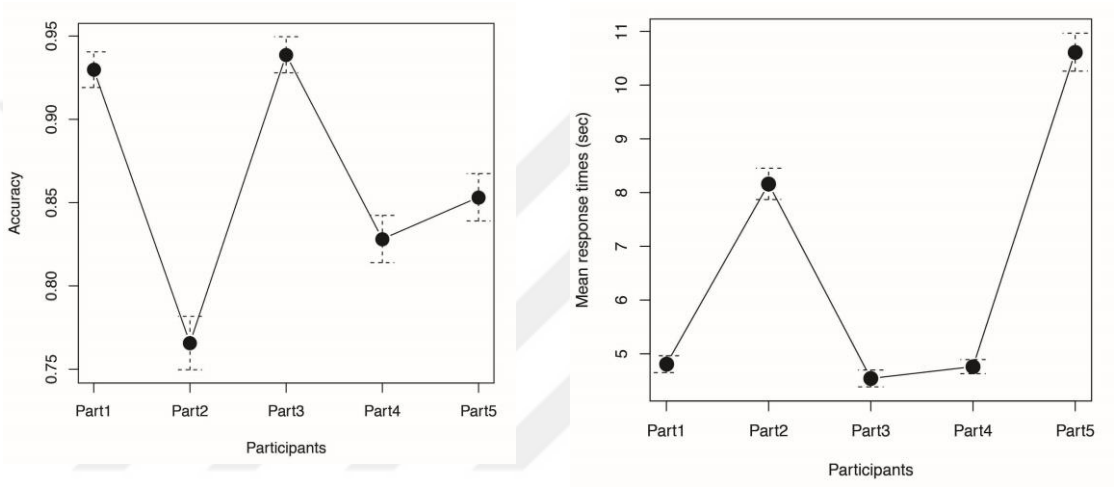

Figure 3.1 : The performance of the participants in terms of accuracy and mean RT. Standard error bars represent ± standard error of the mean.

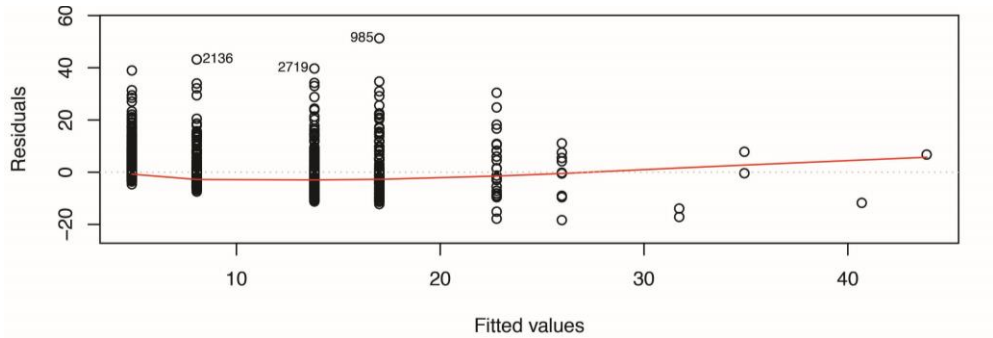

Figure 3.2 : Residuals plots against the model fit the data.

3.2.3 The effectiveness of difficulty levels

Figure 3.3 shows that as the difficulty level gets higher, the students make more mistakes and respond slower. These findings support that our selection of the difficulty levels was right, and the images used for these levels were appropriate. Hence, the color is the simplest attribute to discriminate the objects while the function attribute is

the hardest one. Compared to levels (L1-L3), we noted that all participants performed worse and responded more slowly at L4 (Figure 3.3). Thus, they should be taught more about the function of the objects. We highly recommend students' educators to make a link between the objects and their use in the student's life. Since the drop in accuracy and RT is more from L3 and L4 than from L1 to L2 and L2 to L3, we conclude that the relative difficulty of level L4 respect to level L3 is more than the relative difficulty of level L2 to L1 and level L3 to L2.

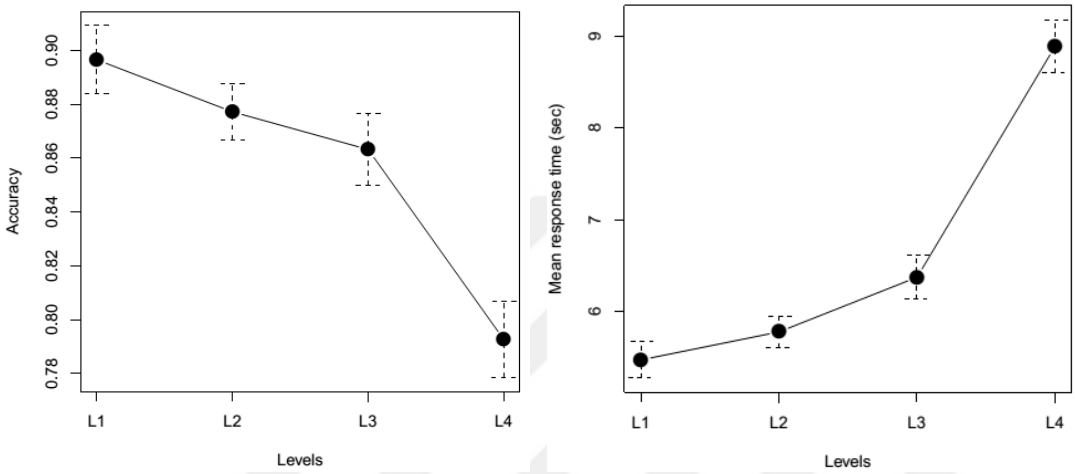

Figure 3.3 : The performance of the difficulty levels in terms of accuracy and mean RT. Standard error bars represent ± standard error of the mean.

3.2.4 The effect of the negative responses

The results in Table 3.1 and Figure 3.4 show that the mean RT for incorrect response trials was always higher than the mean RT for correct response trials. We concluded that the positive responses were faster than the negative responses. There are two justifications to explain this finding: First, our application rewarded the correct responses, and therefore accuracy is emphasized (Ratcliff and Rouder, 1998). Second, in case of incorrect response, the entire memory set must be scanned while the student evaluates the trial (Cowan, 1997). Table 3.1 provides an indication that the incorrect responses have significantly affected the overall performance of students with ASD.

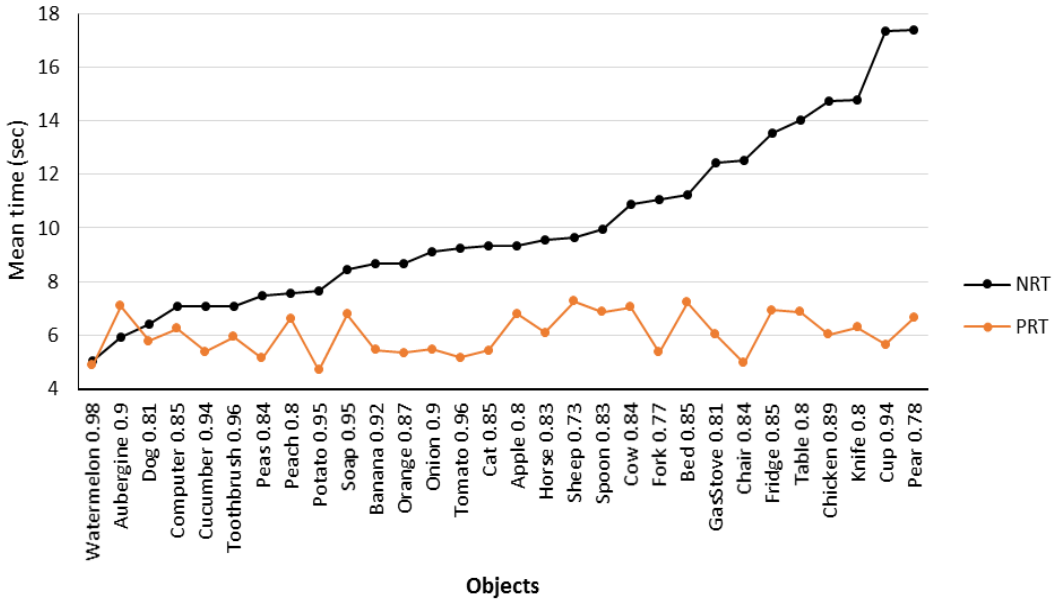

Figure 3.4 : The mean RT for correct and incorrect response trials. The percentage beside the object's name indicates to the accuracy of that object.

3.2.5 The effect of the location of object's image

The location of object's image in a trial was classified into a set of four directions (UpRight, UpLeft, DownRight, and DownLeft). We used one-way ANOVA to test the significant differences in categorical and continuous variables. The significant main effect of the chosen location in trials, $F(3, 308) = 377.8$; $p < .0001$, demonstrated that the location manipulations had a differential influence on the response correct. However, the location manipulations did not influence the correct response time, $F(2, 26) = 0.212$; $p = 0.809$. Thus, the location of object's image in a trial should be uniformly distributed because the location has affected the student's responses. Analyzing the chosen location results, we noted that the students tended to select the location DownRight for negative responses more than other options. This finding was confirmed by the largest number of errors at this location (see Figure 3.5 left). In addition, the location DownRight got a higher percentage of AR (see Figure 3.5 right). This was not sufficient to decide about the way the student followed the images' locations. Future research may include eye trackers to study eye movements during the trial.

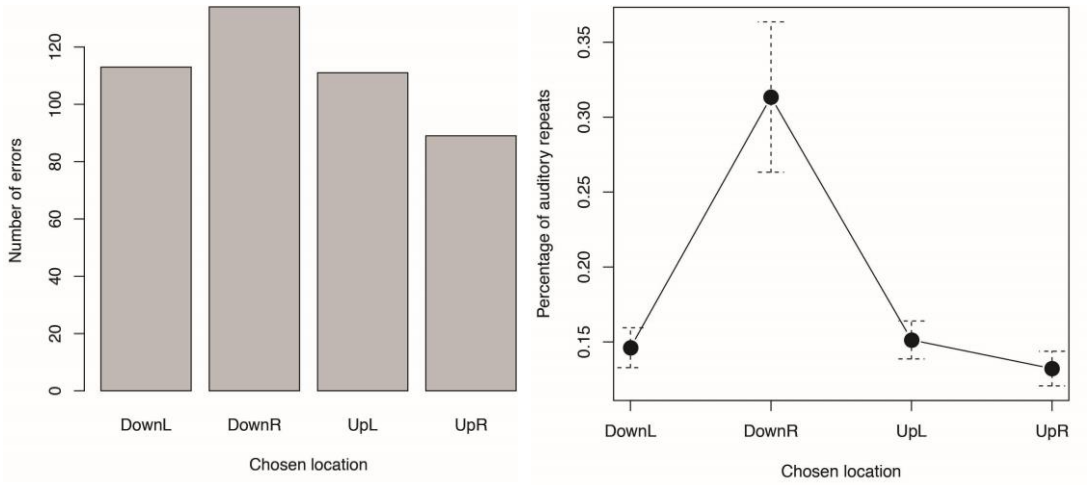

Figure 3.5 : (Left) Chosen location vs. number of errors. (Right) Chosen location vs. percentage of auditory repeats.

3.2.6 The familiarity of categories

The mean performance of all categories is presented in Table 3.2. One-way ANOVA methods was used to assess whether differences existed in the accuracy and RT scores across the five groups of categories indicated a statistically significant difference. There were significant differences across category groups in accuracy means ($F(4,31) = 5.337$; $p < .0002$). Results also evidenced a significant difference between means of category's RTs as determined by one-way ANOVA ($F(4,31) = 4.986$; $p < .0005$). Since some categories were better in accuracy and the others were better in response time, IES was used to determine the categories with better performance. IES scores in Table 3.2 demonstrated that category “vegetables” was the most familiar to our sample of students, while the “animals” and “furniture” categories were the least familiar. The tools and fruits were in the middle with approximately equal IES scores. The reason is that the students see these categories (vegetables, fruits and tools) as an integral part of their daily life, and so they become more familiar than the other categories. However, the number of participants is too small to make strong conclusions about the familiarity of categories. The confusion matrices were created between the true objects and the learned objects in each category for the students’ responses (see Figure 3.6). All correct answers are located in the diagonal of the matrix. Most of the off-diagonal elements of confusion matrix are either zero or have considerably lower values compared to the diagonal elements. A confusion matrix for “vegetables” category was with fewer similarities off-diagonal entries while the similarities were more with higher values for other categories. For animals, the students usually learn animal’s

category by means of pictures not in real world. It is difficult for students to learn about the animals this way. Moreover, learning animals through pictures may influence student's conceptual knowledge of animals (Ganea et al, 2014). Concerning furniture, they have different shapes, patterns, sizes and colors. Due to multiple differences, the students may be confused about how to choose the correct answer.

Vegetables

	Aubergine	Cucumber	Onion	Peas	Potato	Tomato
Aubergine	0.89	0.02	0.04	0.03	0.00	0.03
Cucumber	0.00	0.94	0.02	0.02	0.00	0.01
Onion	0.02	0.03	0.90	0.02	0.02	0.00
Peas	0.04	0.05	0.04	0.83	0.01	0.03
Potato	0.00	0.03	0.00	0.02	0.95	0.00
Tomato	0.01	0.00	0.02	0.00	0.00	0.96

Fruits

	Apple	Banana	Orange	Peach	Pear	Watermelon
Apple	0.80	0.02	0.03	0.07	0.05	0.03
Banana	0.01	0.92	0.02	0.01	0.03	0.01
Orange	0.03	0.02	0.87	0.04	0.03	0.01
Peach	0.05	0.01	0.07	0.80	0.05	0.01
Pear	0.06	0.04	0.03	0.06	0.78	0.03
Watermelon	0.00	0.00	0.01	0.01	0.00	0.98

Animals

	Cat	Chicken	Cow	Dog	Horse	Sheep
Cat	0.85	0.04	0.02	0.08	0.00	0.01
Chicken	0.00	0.89	0.03	0.03	0.02	0.04
Cow	0.04	0.01	0.84	0.04	0.03	0.03
Dog	0.03	0.03	0.08	0.81	0.02	0.03
Horse	0.03	0.05	0.02	0.03	0.84	0.02
Sheep	0.03	0.07	0.13	0.04	0.00	0.73

Furniture

	Bed	Chair	Computer	Fridge	GasStove	Table
Bed	0.85	0.05	0.01	0.04	0.04	0.02
Chair	0.03	0.84	0.03	0.05	0.02	0.03
Computer	0.03	0.03	0.85	0.03	0.05	0.01
Fridge	0.01	0.05	0.01	0.85	0.07	0.02
GasStove	0.02	0.07	0.06	0.03	0.81	0.02
Table	0.04	0.04	0.02	0.04	0.06	0.80

Tools

	Cup	Fork	Knife	Soap	Spoon	Toothbrush
Cup	0.94	0.01	0.01	0.01	0.01	0.01
Fork	0.04	0.77	0.07	0.03	0.05	0.03
Knife	0.01	0.06	0.80	0.03	0.07	0.03
Soap	0.01	0.01	0.01	0.96	0.00	0.01
Spoon	0.04	0.06	0.03	0.01	0.83	0.03
Toothbrush	0.00	0.01	0.01	0.01	0.00	0.96

Figure 3.6 : Confusion matrix for each category. The higher the diagonal entity's value is, the better the performance looks.

Table 3.2 : The overall performance of categories.

Category	Accuracy (%)	RT (sec)	PRT (sec)	NRT (sec)	IES (sec)	AR (%)
Animals	0.822 (0.016)	6.881 (0.267)	6.256	9.775	7.606	0.143
Fruits	0.853 (0.014)	6.665 (0.251)	5.951	10.809	6.976	0.184
Furniture	0.832 (0.013)	7.320 (0.297)	6.393	11.914	7.683	0.183
Tools	0.865 (0.014)	6.705 (0.265)	5.894	11.909	6.813	0.128
Vegetables	0.908 (0.012)	5.713 (0.195)	5.533	7.487	6.096	0.108

3.2.7 Difficulty of recognizing the objects

Figure 3.7 shows that IES for the objects including all levels, and a lower IES score means better performance. We concluded that the five objects that can be easily recognized: potato, watermelon, tomato, cucumber and banana. These objects have very small similarities with values less than 0.03 (see Figure 3.6). We noted that all of them are edible and have distinctive color or shape. The five most difficult objects are sheep, table, pear, apple and bed. It is not usual for the students to see the sheep in their daily life whereas the cat and chicken can be seen in their real life. Pear, peach and apple have high similarities in shape and color. Sheep and cow also have high similarity in shape (see Figure 3.6). The similarities between objects help us how select the distractors in assessment trials for next test sessions. We measure the similarity between objects in a category and conclude how the child confuses them.

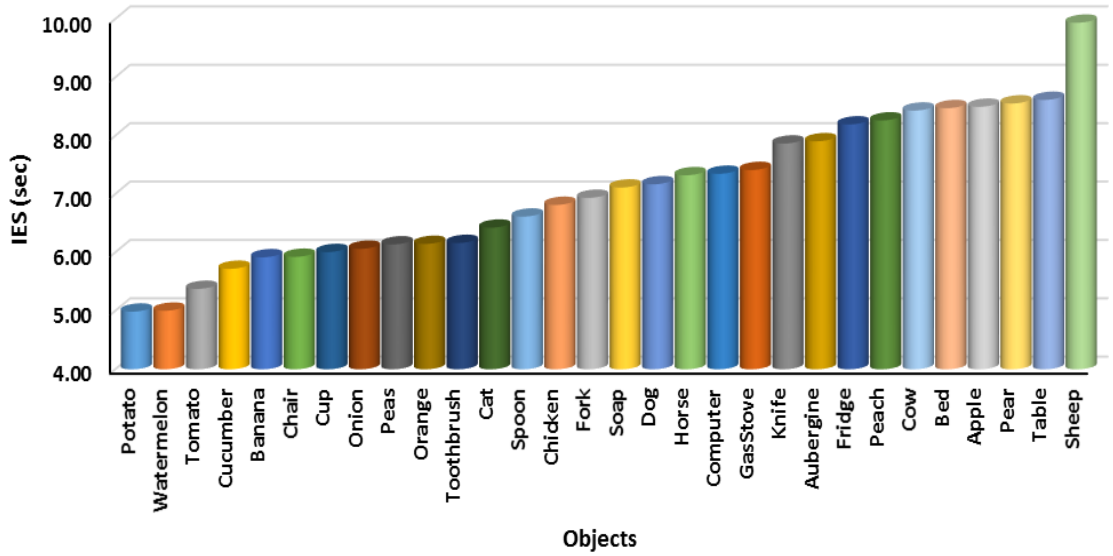

Figure 3.7 : Objects vs. inverse efficiency score (IES). Lower IES scores represent better performance.

3.2.8 Guidance for developing tablet applications

This investigation could not be implemented using traditional approaches for teaching. Using the tablet helped us provide a large number of pictures while teaching. These pictures were presented to students in an easy and interesting way. Tablets also support web-based software and touch screen offers multiple advantages to students. Moreover, the accuracy of the collected data could not be achieved through a paper-based method. Due to these features of tablet PCs, we suggest that tablets can be an effective instructional tool to enhance teaching and learning for students with ASD. Based on our findings, we offer suggestions for developing tablet applications to teach students with ASD. The following should be considered in designing future applications:

- The screen should focus on the objects' images in a trial and should not include many large-size icons or buttons in order to maintain the student's attention.
- When asking about a particular object in a trial, the other three objects should be selected from the same category.
- Distribution of the location of object's image for a trial should be uniform.
- It is possible to stop and pause the application during the session when the student get tired or bored.
- The images of a trial should be uniform in resolution and size.
- The investigation may start with a low-level difficulty and work toward a high-level difficulty.
- The images used for teaching and assessment conditions of a session should be different.
- The participants might either be required to have a set of prerequisite fine tablet skills or be provided with an initial independent set of sessions to increase the tablet usage skills

4. CLASS IMBALANCE AND CLASS NOISE

4.1 Background

Real-world data for many applications including education is never perfect. The data used to make predictions are often imbalanced and suffer from noise. Class imbalance is considered to be one of the ten challenging research problems in data mining (Yang and Wu, 2006). A dataset is said to be imbalanced when one class (minority class) has much fewer instances than the remaining classes (majority classes). The minority class is usually the most interesting with respect to the domain of study; hence high prediction of the minority class is our target. The traditional classification algorithms are designed to maximize the overall accuracy, which is independent of class distribution. Thus, when learning from imbalanced data, they are usually overwhelmed by the majority class instances (Thai-Nghe et al, 2010). "This causes classifiers to tend to overfit and to perform poorly in particular on the minority class" (He and Garcia, 2009). The main difficulty of imbalanced datasets is that a standard classifier might ignore the importance of the minority class because its representation within the dataset is not strong enough and the classifier is biased toward the majority class or, in other words, it is oriented to achieve a good total classification accuracy. Consequently, the instances that belong to the minority class are misclassified more often than those belonging to the majority class. A number of approaches have been proposed to handle the problem of imbalanced classification, both for standard learning algorithms and for ensemble techniques (Guo et al, 2008; Seiffert et al, 2010; López et al, 2013; Zhang et al, 2014; Sluban et al, 2014; Sáez et al, 2015).

The term "noise" refers to data points, which could be considered as erroneous, irrelevant or meaningless. Noise is often divided into two categories (Zhu et al, 2003; Wu and Zhu, 2008): (a) attribute noise (errors or missing values in one or more attributes); and (b) class noise which can be found in the following forms: (1) contradictory instances; instances with the same values of attributes but with different labels. (2) misclassifications; instances with wrong class labels (Zhu et al, 2003).

Regardless of the type of noise, the existence of noisy instances in a dataset has a negative effect on the quality of information retrieved from the data, models built using this data, and decisions made based on the analysis of those models (Zhu and Wu, 2004; Sluban et al, 2014). Noise filtering approaches are often used in machine learning to detect and eliminate noisy instances from training data, and consequently improve the classification accuracy of models induced from the clean data (Zhu et al, 2003; Khoshgoftaar and Rebours, 2007; Anyfantis et al, 2007).

Napierała et al. (2010) categorized the instances for binary problems into three distinct groups: safe, borderline and noisy (see Figure 4.1). Safe instances are placed in relatively homogeneous areas with respect to the class label. Most of the classifiers has capability to correctly identify these instances. Borderline instances are located in the area surrounding decision boundary separating the two classes, where the minority and majority instances may overlap. Finally, noisy instances are those from one class located in the safe areas of the other class. Noisy instance is usually surrounded by instances of an another class. In some studies, noisy instances are referred to as outliers.

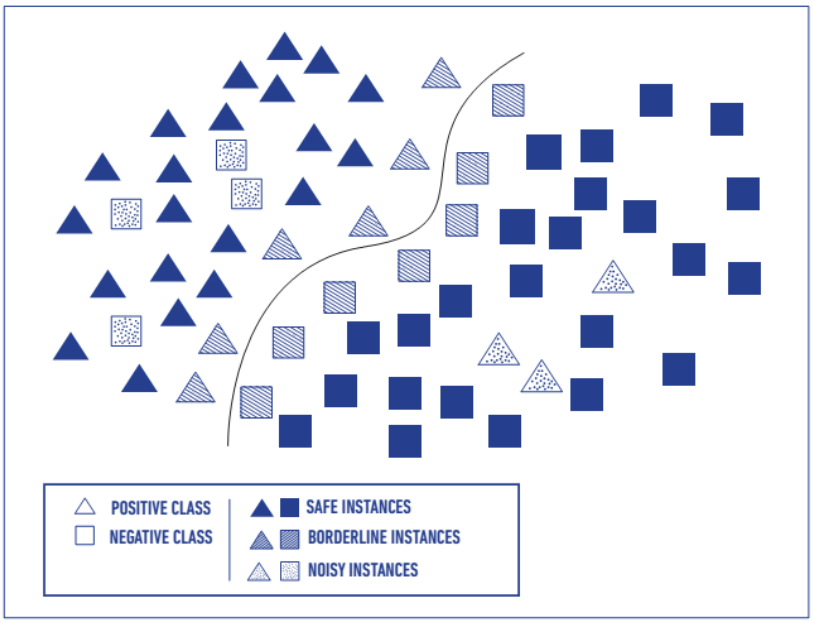

Figure 4.1 : The three types of instances: safe, borderline and noisy instances. The continuous line shows the decision boundary.

In order to achieve better prediction, most of the classification algorithms attempt to learn the borderline instances of each class during the training process (Sáez et al, 2014; Seiffert et al, 2014; Sáez et al, 2015). Napierała et al. (2010) showed that a large

number of borderline instances negatively affected the performance of a classifier. These instances are more likely to be misclassified than the ones located far from the decision boundary, and thus detecting and eliminating noisy instances within borderline area increase the chance to be more effective on the classification results (Sáez et al, 2015).

4.2 Class Imbalance Approaches

Imbalanced data learning is one of the challenging problems in ML. To deal with class imbalance problem (Guo et al, 2008; López et al, 2013), a large number of solutions have been previously proposed. The researchers divided them into data and algorithmic levels. The approaches at data level modify the distribution of the imbalanced dataset, and then the balanced datasets are provided to the learner to improve the detection rate of the minority class. However, the approaches at algorithmic level try to change the search techniques or the classification decision strategies to impose bias toward the minority class or to improve the prediction performance by adjusting weights for each class. Here we summarize some common techniques that are usually used to tackle the class imbalance problem.

4.2.1 Data re-sampling methods

Re-sampling methods aim to balance the class distribution in the training data by either duplicating or generate new minority instances (over-sampling) or removing instances from the majority class (under-sampling) (Thai-Nghe et al, 2010; Blagus and Lusa, 2013). Several techniques for performing over-sampling and under-sampling have been proposed. Both under-sampling and over-sampling have their benefits and drawbacks (Seiffert et al, 2010). While under-sampling is the loss of information that comes from deleting instances, over-sampling could lead to over-fitting that comes from duplicating instances or creating new ones.

4.2.1.1 Random over-sampling

Random over-sampling (ROS) (He and Garcia, 2009; Thai-Nghe et al, 2010) method is used to balance class distribution by randomly duplicating the minority class instances while training the classifier until the desired class ratio is achieved. Sample weight parameter is used in the fit method (Batista et al, 2004). Although ROS is

effective, it may increase the likelihood of over-fitting since it makes exact copies of the minority class instances (Chawla et al, 2002; Guo et al, 2008).

4.2.1.2 Random under-sampling

Random under-sampling (RUS) is also a non-heuristic method that aims to balance class distribution through the random elimination of majority class instances (Batista et al, 2004). The major disadvantage of RUS is that potentially useful instances may be discarded.

4.2.1.3 SMOTE

The Synthetic Minority Oversampling Technique (SMOTE), introduced by Chawla et al. (2002), is one of the most well-known and widely used re-sampling methods. SMOTE generates new artificial minority instances by interpolating among the existing minority instances. Creating "synthetic" instances approach is inspired by a technique that used in handwritten character recognition. Extra training data are created by applying some operations such as rotation and skew on real data (Chawla et al. 2002). This method first finds the k nearest neighbors of each minority instance; next, it selects a random nearest neighbor. Then a new minority class instance is created along the line segment joining a minority class instance and its nearest neighbor. This procedure is repeated until both classes have equal number of instances. SMOTE generates the same number of synthetic data samples for each original minority instance.

4.2.2 Cost-sensitive learning

Standard classifiers assume that the misclassification costs (false negative and false positive cost) are the same for all classes. However, in most real-world applications, this assumption is not true. When learning from imbalanced data, the classifier tends to be biased towards the majority class. Thus, we need to assign a high cost to misclassification of the minority class and try to minimize the overall cost. In the weighting method (Ting, 1998), we assign a certain weight to each instance in terms of its class, according to the misclassification costs, with the minority class given larger weight. Classes with higher weights are given more importance while training the classifier.

4.2.3 Thresholding (cut-off adjustment)

Some classifiers (including random forests) can produce probability estimates on instances, for example: given an instance x, there is 30% probability that it belongs to the positive class and 70% probability that it belongs to the negative class. However, when the classes in the training data are imbalanced, these predictions calculated by the classifier can be inaccurate because many classifiers do not know how to adjust for the class imbalance. If a classifier's probabilities are accurate, the appropriate way to convert its probabilities into predictions is to cut-off at a threshold (usually 0.5) and predict the positive class if the probability is above the cut-off and otherwise the negative class. When the probabilities are inaccurate, this method does not work well. We can improve the predictions by adjusting the threshold to a value that minimizes the total misclassification cost on the training instances, and use this value to predict the class label of test instances (Sheng and Ling, 2006; Blagus and Lusa, 2013).

4.3 Evaluation Metrics in Imbalanced Domains

Evaluation measures are needed so that different classifiers' performances can be compared to each other. For a two-class problem, the confusion matrix, shown in Table 4.1, illustrates the distribution of correct and incorrect instances for the positive and negative classes. TP and TN denote the number of positive and negative instances that are classified correctly, while FN and FP denote the number of misclassified positive and negative instances respectively.

Table 4.1 : Confusion matrix for two-class problem.

	Predicted Positive	Predicted Negative
Actual positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

A number of widely used metrics to measure the performance of the models can be computed based on the confusion matrix. We can obtain the following four metrics from Table 4.1 to measure the classification performance of both, positive and negative, classes independently:

• True positive rate: $TP_{rate} = \frac{TP}{TP + FN}$ is the percentage of positive instances correctly classified as belonging to the positive class.

• True negative rate: $TN_{rate} = \frac{TN}{FP + TN}$ is the percentage of negative instances correctly classified as belonging to the negative class.

• False positive rate: $FP_{rate} = \frac{FP}{FP + TN}$ is the percentage of negative instances misclassified as belonging to the positive class.

• False negative rate: $FN_{rate} = \frac{FN}{TP + FN}$ is the percentage of positive instances misclassified as belonging to the negative class.

Predictive accuracy $Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$ is the performance measure generally associated with classification algorithms and calculated as:

Typically, the accuracy rate is the most commonly used empirical measure, however, since it does not distinguish between the number of correctly classified instances of different classes, accuracy is no longer a proper measure when the data are imbalanced. For example, suppose the dataset has 990 negative instances and only 10 positive instances. If all instances are classified to belong to the majority class, we get 99% accuracy. This result has no meaning because all the positive instances are classified incorrectly (Thai-Nghe et al, 2009). The ROC (Receiver Operating Characteristics) curve is a standard technique for summarizing classifier performance on imbalanced datasets (Fawcett, 2006). The ROC is a graphical plot of the TP_{rate} (sensitivity) against the FP_{rate} (1-specificity) along different threshold values characterizing the overall performance of a studied classifier. The area under the ROC curve (AUC) is the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). AUC values range from 0 to 1, with a higher AUC value indicating that the classifier has a higher discriminative capability to differentiate positive samples from negative samples. An AUC equal to 1.0 indicates a perfect classifier, whereas 0.5 indicates that a model performs like a random classifier (Berrar and Flach, 2011). AUC has been shown to be a reliable measure for imbalanced datasets (García et al, 2012; López et al, 2013; Sáez et al, 2015). Thus, we have used AUC metric to evaluate our experiments in Chapter 5.

4.4 Noise Detection Approaches

Most of data cleaning methods focus on the detection and removal of noise that is the result of an imperfect data collection process. Noise handling is a central focus in many areas of ML. The preliminary approaches to noise treatment aim to create inductive learning algorithms that are able to resist the dataset's noise. However, the most common approaches to noise handling is to eliminate noise by filtering of noisy instances before the model is built. Noise filters are preprocessing mechanisms designed to enhance the data quality by eliminating noisy instances in the training set (Zhu et al, 2003). The separation of noise detection has the advantage that noisy instances do not influence the classifier design. Then a classification algorithm is used on the reduced and clean training data.

There are numerous noise detection approaches in literature. The most widely used are Classification Filter (CF) proposed by Gamberger et al. (1999), Ensemble Filter (EF) (Khoshgoftaar et al, 2006), and Iterative-Partitioning Filter (IPF) (Zhu et al, 2003). CF approach identifies the noisy instances using cross validation. The training set is partitioned into $\boldsymbol{n}$ subsets, then $\boldsymbol{n}$ classifiers are trained using an aggregation of any $\boldsymbol{n-1}$ subsets to identify noise from the complementary (excluded) subset. On the other hand, the EF approach learns a classifier from each single subset and then the classifier is evaluated on the whole dataset. To identify noisy instances, CF fully trusts $\boldsymbol{n}$ classifiers, i.e., the instance incorrectly classified by one classifier is directly identified as noise, but EF takes the majority or consensus vote of a committee of $\boldsymbol{n}$ classifiers (Zhu et al, 2003). IPF is a preprocessing technique based on the EF. It removes noisy instances in multiple iterations until the number of identified noisy instances in each of these iterations is less than a certain percentage of the size of the original training data (Sáez et al, 2015). Noise Filters have two major drawbacks: some removed instances from the dataset are valuable, and a number of classifiers have to be trained so that filtering can be performed, which could be time consuming.

4.5 Machine Learning for student performance prediction

In recent years, there has been an increased interest in data mining and ML to answer research questions for educational purposes. This new emerging area of study is called educational data mining (EDM) (Romero and Ventura, 2007) which can be effectively

applied in the field of education (Márquez-Vera et al, 2013; Osmanbegović et al, 2015; Satyanarayana and Nuckowski, 2016). EDM is concerned with "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist" (Romero & Ventura, 2013, p. 12). Predicting student's performance is one of the useful applications of EDM and its aim is to estimate the values of an unknown output (response or grade) based on the predicting attributes. In EDM, outlier and noise analysis can be used to improve data understanding, and detect students with learning problems (Romero and Ventura, 2007).

Researchers in the educational community have been interested in applying ML techniques for predicting student performance. Class imbalance problem was one of the obstacles that caused unsatisfactory prediction results. Thai-Nghe et al. (2009) proposed a model to improve the student academic performance prediction by dealing with the class imbalance problem. They first re-balanced the dataset by SMOTE and then used cost-insensitive learning to minimize the misclassification cost. The model was examined on four datasets, and the results improved compared to the baseline classifiers. Márquez-Vera et al. (2013) devised a genetic programming algorithm with a data balancing approach for solving the problem of student failure using real data about 670 first-year high school students in Mexico. Chau and Phung (2013) proposed an approach with a hybrid re-sampling scheme and random forest for student's final status prediction in an academic credit system. The proposed approach can deal with class imbalance issues and missing data. In a more recent study, an ensemble filtering approach has been used by Satyanarayana and Nuckowski (2016) to improve the quality of student data by eliminating noisy instances. They showed that, compared to single filters, using ensemble filters gives better predictive accuracies on student performance data.

4.6 Literature Review

With reference to the original SMOTE technique, several adaptations have been proposed in the literature (He and Garcia, 2009; Sáez et al, 2015), most of them aim at identifying the region in which the minority instances should be generated. Another extension of SMOTE corresponds to the Borderline-SMOTE technique (Han et al,

2005), which only over-sampled the minority instances near the borderline since these are more likely to be misclassified. This method achieves better TP rate and F-value than SMOTE and random over-sampling methods. In (García et al, 2012) three approaches were proposed to modify original SMOTE based upon the concept of surrounding neighborhood. The surrounding SMOTE methods generate artificial minority instances but taking into account both the proximity and the spatial distribution of the instances in order to extend the regions of the minority class.

Batista et al. (2004) proposed two data cleaning methods to the over-sampled training set by SMOTE. In the first method, SMOTE-TL (SMOTE with Tomek links) uses Tomek links to remove instances after applying SMOTE. An instance is removed either because it is noisy or because it is near the border. Tomek links (Tomek, 1976) are defined on pairs of minimally distanced nearest neighbors of opposite classes. Let S_{min} and S_{maj} denote the minority and majority classes respectively. Given an instance pair (x, y), where $x \in S_{min}, y \in S_{maj}$ and $d(x, y)$ the distance between x and y, then (x, y) pair is called a Tomek link if there is no instance z such $d(x, z) < d(x, y)$ or $d(y, z) < d(x, y)$. SMOTE-TL is used as an under-sampling method, so only majority examples are removed (Batista et al, 2004). The second method SMOTE-ENN (SMOTE with Edited Nearest Neighbor) tends to remove more instances (from both classes) than the SMOTE-TL does, so it is expected to provide a more effective data cleaning. ENN removes from the training set any instance that differs from two of its three nearest neighbors (Batista et al, 2004). The two methods SMOTE-TL and SMOTE-ENN are characterized by their over-sampling followed by under-sampling.

In ML, ensemble methods combine several individual classifiers to construct a new classifier that is (often) more accurate and reliable than any of its member classifiers (Dietterich, 2000). Researchers have used ensemble methods to deal with the problem of noise filtering. Consensus and majority are the two voting schemes that could be implemented to identify noisy instances. The former eliminates an instance if it is misclassified by all the classifiers, while the latter eliminates an instance if it is misclassified by more than half of the classifiers (Sáez et al, 2015). There are many ensemble-based noise filters. For example, Brodley and Friedl (1999) uses consensus filters and majority vote filters to identify and eliminate mislabeled training instances, which are incorrectly classified by the multiple classifiers. The results show that if that

the training data set is sufficiently large, then classification accuracy can be improved as more noisy instanced were removed. Sluban et al. (2014) presented an ensemble-based methodology for explicit noise detection and ranking, called NoiseRank. It can rank the detected noisy instances according to the predictions of several different noise detection algorithms, and thus it provided more reliable results. NoiseRank was successfully applied to a real-life coronary heart disease (CHD) patient data.

Authors in (Seiffert et al, 2014) performed a comprehensive and empirical study on the effects of class imbalance and class noise on 11 different classification algorithms and data sampling techniques when they used to predict software quality. They compared the performance of seven sampling techniques using 12 datasets derived from real world software quality data with different levels of class noise and imbalance. Later, Sáez et al. (2015) proposed and examined a new extension of SMOTE through an IPF filter, called SMOTE-IPF, which can overcome the problems introduced by noisy and borderline instances in imbalanced datasets. The experiments were carried out both on a set of synthetic datasets with different levels of noise and shapes of borderline examples as well as real-world datasets. The results show that SMOTE–IPF performed better than existing SMOTE generalizations with both synthetic and real-world datasets. It also outperformed the rest of the methods with the real-world datasets with additional noise.

5. LEARNING PERFORMANCE PREDICTION ON EDUCATION DATA

A decision making program is typically called training, where collected instances or input data are called the training set, and the program is referred to as a model or learning algorithm. Each instance in the training set has a class label such as spam/not-spam or success/fail. The instances in unseen data (called testing set) do not have labels. The input data are represented by a set of features. The first step while building a model is deciding what features of the input are relevant, and how to encode them. The goal is to ”train” a model based on the training data and use this model to predict the class labels for instances in testing set. “The goal of inductive learning algorithms is to form generalizations from a set of training instances such that the classification accuracy is maximized. This maximum accuracy is usually determined by: (1) the quality of the training data; and (2) the inductive bias of the learning algorithm” (Zhu and Wu, 2004, p. 177). The quality of a dataset can be characterized by its features and class labels.

In this Chapter, we carried out a series of experiments to examine the impact of imbalance, noisy and borderline instances on a set of classifiers. The classifiers are used to predict the performance of students with ASD whilst they have been teaching object recognition. The dataset used to make the prediction will be described in Section 5.2. We claim that the distribution of borderline instances and the existence of noise in a training data brings about difficulties ML algorithms, and thus the identification and elimination partially or completely of noise should improve the classifier’s effectiveness. However, in some cases, the amount of noise can be large. In this work, we consider only noisy instances restricted within the area surrounding class boundaries. Thus the noisy instances, located inside the borderline area, from the majority class are removed and the minority class remains unchanged. The idea behind this is to keep the instances from minority class. We propose two noise filters to eliminate the noisy instances from the imbalanced dataset.

5.1 Proposed Methods

We present two empirical methods that address noise filtering and class imbalance problems simultaneously, the first is class-Balanced by SMOTE & Thresholding combined with Classification Filter, called BST-CF, and the second is class-Balanced by SMOTE & Thresholding combined with Ensemble Filter, called BST-EF. Both methods eliminate noisy instances, located inside the borderline area, from the majority class and the minority class remains unchanged. Our methods incorporate a noise filtering into two class-imbalance techniques SMOTE and thresholding which are used to balance the class distribution of the training data and choose the best boundary between classes. The CF approach is used to identify noisy instances in the first method while the EF approach is used in the second. The main differences between our methods and other noise filters are:

- The noisy instances are refined and restricted within a borderline area of majority class.
- To the best of our knowledge, thresholding together with noise filtering approach has not been used before.

Our proposed method BST-CF is described as follows:

1. We use random forest (Breiman, 2001) as the base classifier. Estimate the optimal model hyperparameters for RF. Grid search with cross-validation is used to determine the best parameters in terms of AUC that can be used to build an accurate RF model.
2. Split the whole dataset X (including baseline + RT features) into five subsets, four of them are selected as training data X_{train}, and the fifth subset is used as the test data X_{test} to evaluate the performance of RF model on the clean data.
3. Initially, the method starts with a set of noisy instances $S = \emptyset$.
4. Use stratified 10-fold cross-validation to divide X_{train} into fit set X_{fit} (aggregation of any 9 subsets) and validation set X_{val} (excluded subset) such that they are randomly sampled, and classes are equally balanced in both.
5. Firstly, oversampling technique SMOTE is applied to balance X_{fit}, then the RF model is built on this set.

6. Investigate the threshold value θ^*_{fit} in range from 0.40 to 0.80 that yields the best predicted probabilities on X_{fit}. The measure AUC is used for estimation of the threshold.
7. Generate the predicted probabilities that the RF classifier assigned to each instance in validation set X_{val}.
8. Find out the borderline instances in X_{val}, i.e., instances around the chosen threshold θ^*_{fit} with a certain β width.
9. Add to set S the noisy instances which are incorrectly classified instances within the borderline area and belonging to the majority class.
10. Repeat steps 4-9 on the other folds, and then eliminate the noisy set S.
11. Re-fit the RF model on the filtered fit set X'_{train}.
12. Find the optimal threshold value θ^*_{val} by taking the mean of θ^*_{fit} of the ten runs.
13. The final RF model generated in step 11 is then used to predict the class label of test data X_{test} using threshold θ^*_{val}.

Evaluation measures are estimated in 10-fold CV repeated 5 times and the results are obtained by averaging scores of the 5 runs. Figure 5.1 illustrates the overall design of our methods.

The second method BST-EF differs from the BST-CF method in the following:

1) The fit set X_{fit} consists of only one subset while the validation set X_{val} is a union of remaining 9 subsets (step 4 above).
2) When identifying a noisy instance, BST-CF adds it directly to S set at each run, but BST-EF computes the number of times that an instance is identified as a noise. Then eliminates the noisy instances with high scores; such that the total number of eliminated noisy instances in S equals the average number of noisy instances identified at each of the ten runs (step 9).
3) The optimal threshold θ^*_{val} is found by cross validation on the set X'_{fit} (step 12).

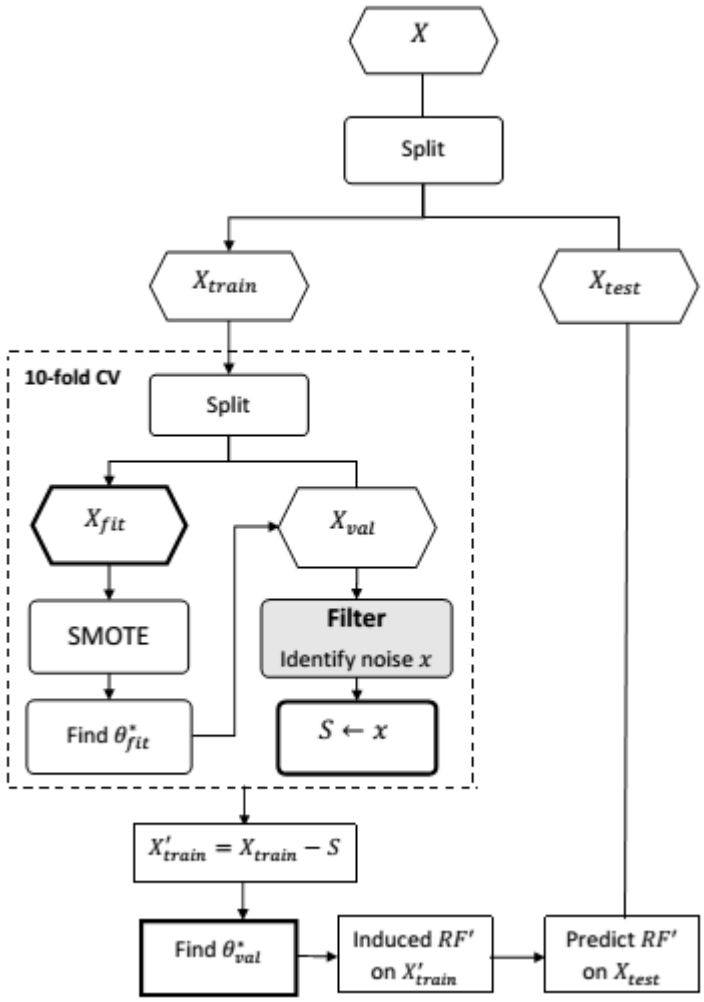

Figure 5.1 : The overall design of BST-CF and BST-EF. The steps that differ between two methods are shown using a thick border.

5.2 Dataset

The dataset used in this chapter has been gathered from the web application during learning sessions of the investigation described in Chapter 2. Each student performed 20 sessions. The total number of instances (rows) used in this study is 3090. Each instance represents the results of an assessment trial for a particular student during the conduction of the investigation. The attributes/variables of the dataset are shown in Table 5.1. Five predictors (student_id, object_id, category_id, level and response time (RT)) and one output variable (response correct (RC)) were used in our analysis. Since they contain discrete and unordered values, 1-of-K representation was used for the following attributes: student_id, object_id, and category_id. After applying this representation, the total number of features was 42.

One of the goals of the study is to investigate if the RT attribute can help improve the RC prediction. Therefore, two feature sets are created. The baseline feature set consists of all the features excluding RT, while the second set is the baseline features with the RT.

Table 5.1 : Description of the dataset used.

Attribute	Description	Data Type	Values
student_id	The student ID	categorical	1 to 5 → binary (5)
object_id	The object ID used in a trial	categorical	1 to 30 → binary (30)
category_id	The category ID of object	categorical	1 to 5 → binary (5)
level	Difficulty level of the object	integer	1 to 4
RT	The response time	real	in seconds
RC	The response by the student	binary	0 or 1

Since we want to predict the RC, we have a binary classification problem. As 85.6% instances are labeled with 1 (correct) and the remaining 14.4% of instances are labeled with 0 (incorrect), this is an imbalanced dataset and the imbalance ratio (IR) of 1 to 0 instances is 5.94. In this study (and without loss of generality), the minority class (class 0) is regarded as the positive class, and the majority class (class 1) is regarded as the negative class.

Our learning dataset contains, by its nature, many instances with similar attribute values especially in the baseline feature set. This is because each object from the same level was shown at least two times in assessment trials during 20 sessions for each student. However, including all features, there are 267 duplicate instances; only 46 of them are contradictory instances (they have the same attribute values and a different class). The small number of contradictory instances does not significantly affect the performance. The dataset also has a considerable number of mislabeled instances, as we will see in Section 5.3, when different classification algorithms are employed without any additional approaches handling class imbalance. The dataset does not contain attribute noise (with the exception of the RT attribute), and there are no rows with missing values. According to (Zhu and Wu, 2004), it is not recommended to eliminate instances, which contain attribute noise; because the other attributes of those instances may have valuable information. In this study, we focused on class noise caused by mislabeling; on the other hand, we ignored class noise caused by contradictory instances as well as attribute noise.

5.3 Experiments and Results

5.3.1 Experiment 1 – Class Imbalance

In the first set of experiments, we analyzed the behavior of class imbalance approaches. Our aim is to investigate the improvement of classifier's result when using these approaches. Different ML algorithms were built, and their performances were evaluated through the widely used stratified 10-fold cross validation method. In this method, our educational dataset is randomly divided into ten subsets of approximately equal sizes. Nine of these ten subsets are selected as training data, and the tenth subset is used as the test data to evaluate the performance of the trained ML model. This procedure is repeated for 10 times for each model. All experiments have been conducted using scikit-learn and UnbalancedDataset packages. Scikit-learn (Pedregosa et al, 2011) is a general purpose ML library written in Python which provides efficient implementations of many ML algorithms. UnbalancedDataset is a python package that provides a number of implementations of SMOTE as well as various other re-sampling techniques (Url-2, 2016).

Three base ML classification algorithms were used: Logistic Regression (LR), Random Forest (RF) classifier and linear Support vector machine (SVM). The RF involves an ensemble of decision trees grown based a randomly selected subset of samples and features. The prediction is made by aggregating majority vote of the ensemble. For the RF model we set the tree number parameter to 200. Linear SVM maps input data into a high-dimensional feature space with a linear kernel function. Penalized classification imposes an additional cost on the model for making classification mistakes on the minority class during training. To improve scores for LR, AdaBoost was used on the best LR model we had. AdaBoost is an ensemble classifier which follows a boosting technique that combines multiple weak classifiers into a strong one. Scikit-learn's AdaBoost implementation can take a classifier as input (base estimator). To compare results, random classifier (RN) is regarded as a baseline classifier because by intuition any other model should outperform this classifier. We used RN that makes predictions uniformly at random.

Firstly, we built these five classification models using baseline feature set and we got above 88% accuracy by all classifiers except RN (see Table 5.2). Actually this accuracy reporting is misleading on the results, because, as it can be seen in Table 5.2,

the algorithms classified all instances to the majority class. The results indicate that all classifiers have the test AUC score about 0.5 which means that they perform like a random classifier.

Table 5.2 : The results of classification algorithms using baseline feature set.

Algorithm	Accuracy	AUC	Normalized Conf. matrix
RN	0.519	0.489	[0.44 0.56] [0.49 0.51]
LR	0.886	0.495	[0 1.0] [0.01 0.99]
RF	0.887	0.508	[0 1.0] [0.02 0.98]
SVM	0.896	0.500	[0 1.0] [0 1.0]
AdaBoost	0.883	0.489	[0 1.0] [0.01 0.99]

In order to overcome the problem of class imbalance, five different approaches are implemented: ROS, SMOTE, SMOTE-TL, SMOTE-ENN and weighting. Table 5.3 and 5.4 present the overall test AUC results (± for standard deviation) obtained by different classifiers using class-imbalance approaches considered in Chapter 4. The baseline feature set is included in Table 5.3 while all features are included in Table 5.4. The column denoted by 'None' corresponds to the case in which no class-imbalance approach is performed. The best case for each approach is highlighted in bold. From results of Table 5.3 and 5.4, the following main points should be stressed:

1- Classification with balanced dataset results in effective performance, regardless of any used classifier. In most cases, LR outperforms the others classifiers (RF, SVM and AdaBoost with LR).
2- The results indicate that the classifiers did achieve better when RT added to the baseline features, and therefore the RT attribute helped improve the performance prediction.
3- All approaches have been able to improve the AUC results. However, SMOTE-ENN and weighting are quite robust on average among those approaches for all classifiers. Thus, these two approaches will be used in experiment 3.

5.3.2 Experiment 2 – Class Imbalance and Noise Filter

We examine the proposed methods presented in Section 5.1 by increasing the width of borderline for the majority class. Different widths β are used from 0 to 0.6, where BSTCF-0 indicates that BST-CF is applied to the original training data (without any noise elimination) and BSTCF-0.1 indicates the BST-CF with borderline's width equals 0.1. A similar method is used for BST-EF. With the two methods, the parameters involved in RF model were empirically chosen for the best results and described as follows: number of trees in the forest ; the minimum number of samples required to split an internal node; the minimum number of samples in newly created leaves . The RF is run using the following parameters:

Parameters	value
n_estimators	120
min_samples_split	5
min_samples_leaf	6
n_jobs	-1

In this experiment, we compare the results through AUC and TP rate. TP rate reflects the performance of the classifier on the minority class of test set, while AUC shows the performance of the classifier on the whole test set.

Tables 5.5 and 5.6 demonstrate results for BST-CF and BST-EF at different widths on the whole dataset DS. The tables also show the amount and proportion of instances detected as noise by filters at different widths. The bold numbers present the best results among these β. Experimental results have shown that our proposed methods are effective and robust in identifying noise and improving the AUC scores. They achieve considerable improvements with respect to the original SMOTE. In Table 5.4, AUC score for RF with SMOTE is 0.593 whereas BSTCF-0.4 and BSTEF-0.4 increase AUC to be 0.709 and 0.711 respectively. In general, the two methods have slight differences in term of AUC at different widths. However, BST-EF achieves better TP rate than BST-CF. We can also observe that BST-CF is more aggressive in terms of removed noisy instances.

Table 5.3 : Comparison of AUC results for different classifiers using baseline feature set.

Algorithm	None	ROS	SMOTE	SMOTE-TL	SMOTE-ENN	Weighting
RN	0.490 ±.031	0.505 ±0.212	0.508 ± 0.062	0.502 ± 0.033	0.490 ± 0.032	0.495 ± 0.022
LR	0.501 ± .004	0.593 ± 0.070	**0.662** ± 0.041	**0.661** ± 0.046	**0.664** ± 0.050	0.665 ± 0.039
RF	0.500 ± .003	0.622 ± 0.039	0.626 ± 0.025	0.621 ± 0.042	0.632 ± 0.040	**0.678** ± 0.047
SVM	0.5 ± 0.0	0.578 ± 0.045	0.656 ± 0.039	0.656 ± 0.041	0.660 ± 0.049	0.653 ± 0.038
AdaBoost	0.501 ± .004	**0.639** ± 0.047	0.657 ± 0.036	0.659 ± 0.038	0.658 ± 0.043	0.651 ± 0.030

The best case for each approach is highlighted in bold. (± standard deviation of the AUC mean).

Table 5.4 : Comparison of AUC results for different classifiers using baseline with RT features.

Algorithm	None	ROS	SMOTE	SMOTE-TL	SMOTE-ENN	Weighting
RN	0.507 ± .033	0.515 ± 0.048	0.502 ± 0.049	0.489 ± 0.039	0.503 ± 0.039	0.515 ± 0.040
LR	0.534 ± .015	0.643 ± 0.067	**0.682** ± 0.038	**0.687** ± 0.042	0.690 ± 0.027	**0.690** ± 0.039
RF	0.509 ± .010	0.631 ± 0.042	0.593 ± 0.030	0.602 ± 0.039	0.677 ± 0.044	0.682 ± 0.043
SVM	0.5 ± 0.0	**0.649** ± 0.054	0.679 ± 0.047	0.684 ± 0.044	**0.695** ± 0.035	0.681 ± 0.041
AdaBoost	**0.536** ± .024	0.638 ± 0.047	0.678 ± 0.034	0.684 ± 0.039	0.686 ± 0.026	0.678 ± 0.028

The best case for each approach is highlighted in bold. (± standard deviation of the AUC mean).

Table 5.5 : Results for BST-CF with different borderline's width on dataset DS.

Method	AUC	TP_{rate}	TN_{rate}	Count noise	% noise
BSTCF-0.0	0.662 ± 0.027	0.605 ± 0.062	0.716 ± 0.028	0	0
BSTCF-0.1	0.685 ± 0.019	0.696 ± 0.049	0.675 ±0.014	130	5.26
BSTCF-0.2	0.689 ± 0.016	0.723 ± 0.056	0.641 ±0.025	249	10.07
BSTCF-0.3	0.697 ± 0.012	0.752 ± 0.066	0.636 ± 0.033	340	13.75
BSTCF-0.4	**0.709** ± 0.020	0.787 ± 0.020	0.626 ± 0.018	413	16.71
BSTCF-0.5	0.706 ± 0.022	**0.790** ± 0.058	0.615 ± 0.010	484	19.58
BSTCF-0.6	0.705 ± 0.015	0.796 ± 0.049	0.605 ± 0.019	524	21.20

Table 5.6 : Results for BST-EF with different borderline's width on dataset DS.

Method	AUC	TP_{rate}	TN_{rate}	Count noise	% noise
BSTEF-0.0	0.676 ± 0.048	0.775 + 0.065	0.590 ± 0.024	0	0
BSTEF-0.1	0.684 ± 0.020	0.833 ± 0.060	0.535 ± 0.016	133	5.38
BSTEF-0.2	0.695 ± 0.020	0.809 ± 0.071	0.580 ± 0.037	244	9.87
BSTEF-0.3	0.703 ± 0.029	**0.852** ± 0.056	0.532 ± 0.036	330	13.35
BSTEF-0.4	**0.711** ± 0.028	0.820 ± 0.061	0.601 ± 0.040	395	15.98
BSTEF-0.5	0.705 ± 0.025	0.811 ± 0.057	0.619 ± 0.037	437	17.68
BSTEF-0.6	0.693 + 0.028	0.787 + 0.50	0.605 + 0.020	489	19.78

After analyzing the distribution of noisy and eliminated instances in BST-CF and BST-EF, we find out that many of them (about 35%) belong to student2. Thus, the classifier's perdiction is highly affected by student2's data and consequently we excluded them from the whole dataset DS to get a better prediction. Tables 5.7 and 5.8 summarize the results for BST-CF and BST-EF with different $\boldsymbol{\beta}$ on the dataset DS-2 which is produced by excluding student2's data, i.e., student2' sessions' trials were removed from DS. It is clear that applying our methods on DS-2 improve the performance. Moreover, BST-EF achieves a slightly better AUC score than BST-CF.

Table 5.7 : Results for BST-CF with different borderline's width on dataset DS-2.

Method	AUC	TP_{rate}	TN_{rate}	Count noise	% noise
BSTCF-0.0	0.694 ± 0.030	0.632 ± 0.072	0.755 ± 0.035	0	0
BSTCF-0.1	0.711 ± 0.037	0.713 ± 0.080	0.708 ± 0.048	90	4.71
BSTCF-0.2	0.713 ± 0.033	0.731 ± 0.066	0.694 ± 0.044	164	8.58
BSTCF-0.3	0.718 ± 0.028	0.763 ± 0.043	0.674 ± 0.044	218	11.40
BSTCF-0.4	**0.725** ± 0.021	0.784 ± 0.036	0.666 ± 0.043	284	14.85
BSTCF -0.5	0.718 ± 0.019	**0.809** ± 0.041	0.627 ± 0.030	341	17.83
BSTCF -0.6	0.721 ± 0.024	0.795 ± 0.041	0.647 ± 0.037	367	19.19

Table 5.8 : Results for BST-FF with different borderline's width on dataset DS-2.

Method	AUC	TP_{rate}	TN_{rate}	Count noise	% noise
STEF-0.0	0.707 ± 0.032	0.763 ± 0.053	0.660 ± 0.039	0	0
STEF-0.1	0.714 ± 0.014	0.788 ± 0.040	0.641 ± 0.039	92	4.81
STEF-0.2	0.718 ± 0.033	0.798 ± 0.089	0.618 ± 0.053	166	8.68
STEF-0.3	0.722 ± 0.023	0.806 ± 0.046	0.635 ± 0.023	226	11.82
STEF-0.4	0.718 ± 0.019	0.742 ± 0.019	0.683 ± 0.048	267	13.96
STEF-0.5	**0.731** ± 0.017	**0.834** ± 0.055	0.628 ± 0.068	302	15.79
STEF-0.6	0.719 ± 0.016	0.781 ± 0.035	0.657 ± 0.06	319	16.68

The goal of experiment 2 is to investigate the performance of the two methods BST-CF and BST-EF for different classifiers. Thus, we introduce the following hypothesis: the AUC results obtained by the combinations of one of proposed methods and a classifier are significantly better than the AUC results obtained from only a classifier with class-imbalance approach. Based on the results in above Table 5.5 - 5.8, the best AUC scores by the BST-CF and BST-EF methods produced the optimal noise set to be removed from dataset. In this experiment, all the classification algorithms are executed using stratified 10-fold cross validation. The experimental results, shown in Figures 5.2 and 5.3, are presented in terms of the average AUC via 10 runs for different four classifiers after removing noisy instances from dataset DS and DS-2 using BST-CF method. Similarly Figures 5.4 and 5.5 depict the results for the second method BST-EF. These four figures reveal the following conclusions. First of all, the four classifiers improve AUC scores significantly for all five class-imbalance approaches (solid line) compared with results obtained without noise removal like that in experiment 1 (dashed line). Secondly, SMOTE-ENN and weighting are robust on average among other approaches. The observation of the best AUC's scores for BST-CF (0.747 with DS and 0.772 with DS-2) show that they are achieved by RF classifier. Third, the AUC results of BST-CF are slightly better than BST-EF. The reason may be due to BST-CF eliminates more instances. Finally, comparing with the whole data DS, the results obtained by two methods on DS-2 are better than in terms of AUC although the size of the dataset has decreased by 22%.

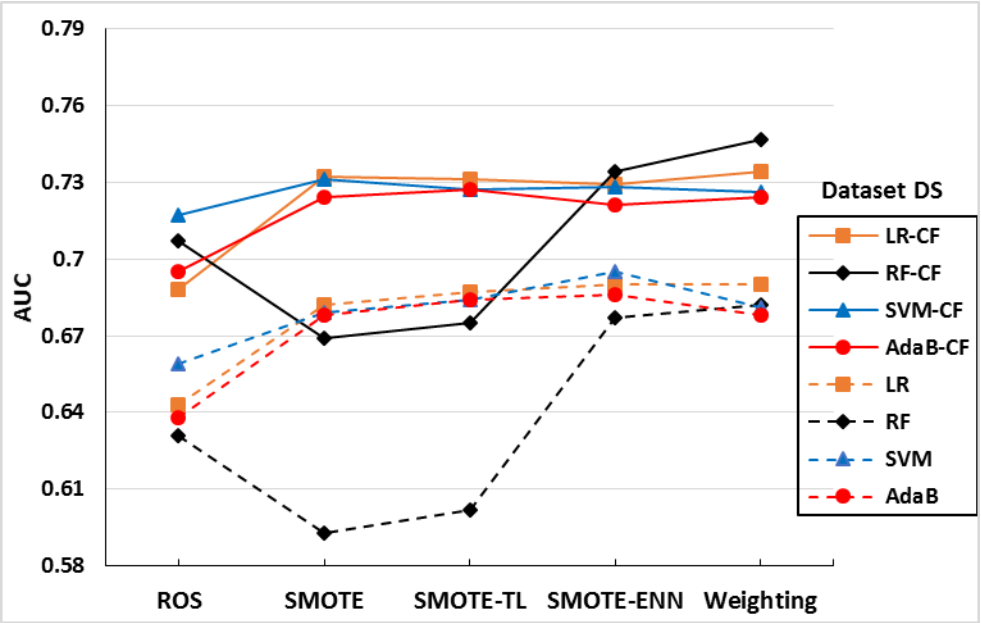

Figure 5.2 : Average AUC obtained by different classifiers using only class-imbalance approaches (dashed line) and with BST-CF (solid line) on dataset DS.

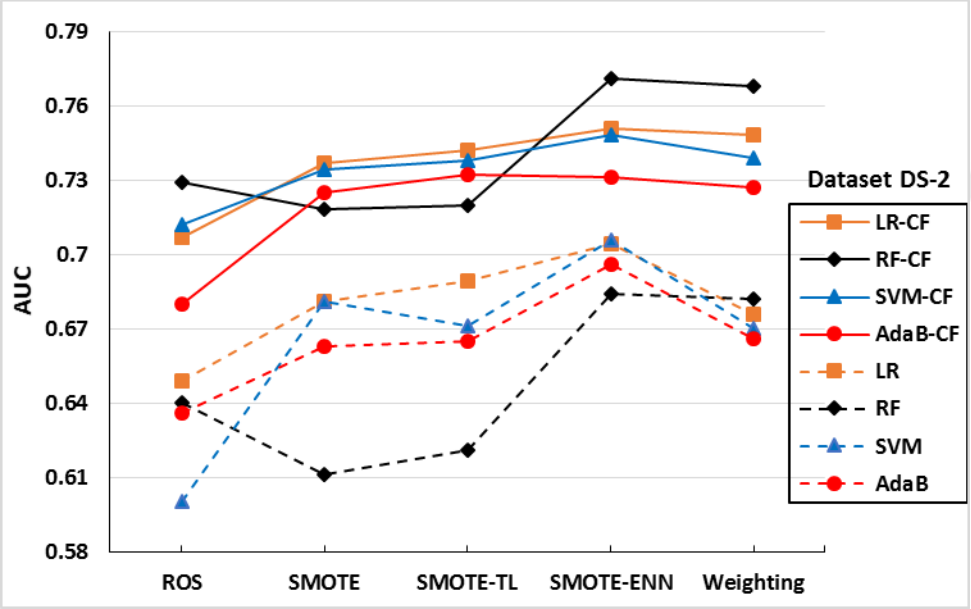

Figure 5.3 : Average AUC obtained by different classifiers using only class-imbalance approaches (dashed line) and with BST-CF (solid line) on dataset DS-2.

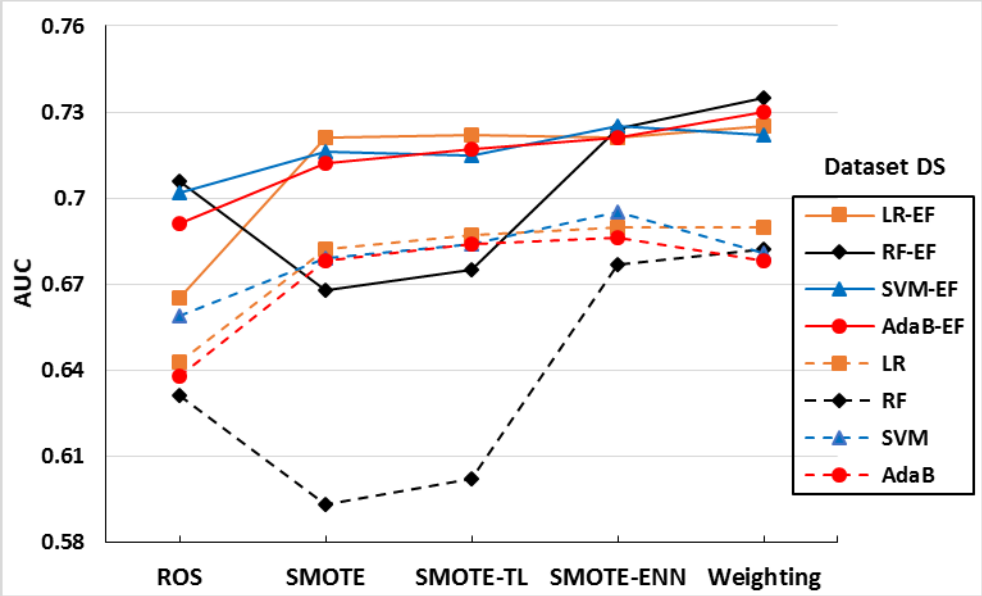

Figure 5.4 : Average AUC obtained by different classifiers using only class-imbalance approaches (dashed line) and with BST-EF (solid line) on dataset DS.

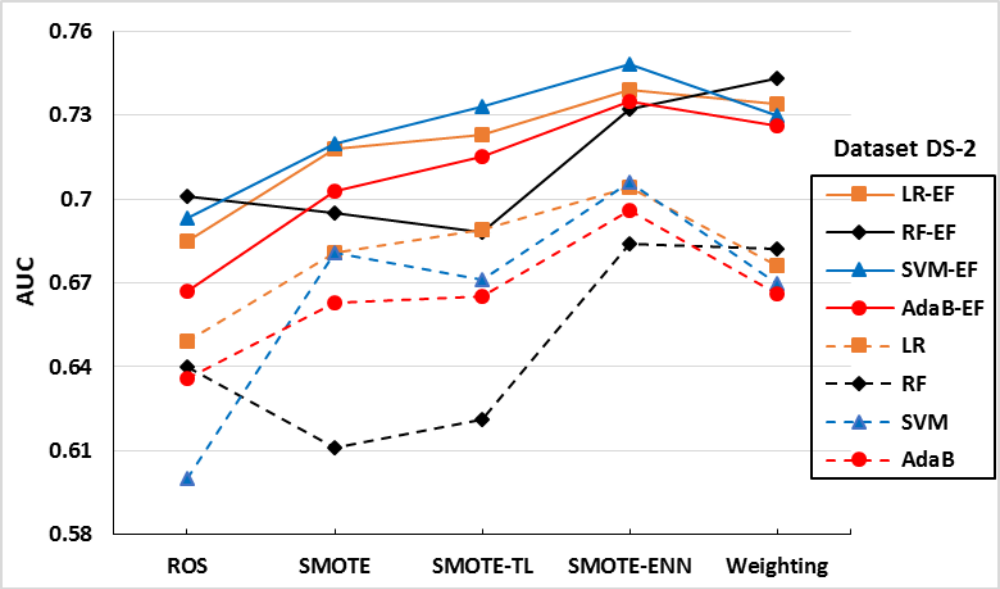

Figure 5.5 : Average AUC obtained by different classifiers using only class-imbalance approaches (dashed line) and with BST-EF (solid line) on dataset DS-2.

The Friedman test (García et al, 2010) is a non-parametric statistical test of the parametric two-way analysis of variance. The objective of this test is to determine if there are differences among treatment effects across multiple test results. Using the results of five class-imbalance approaches employed in this study on two datasets (DS and DS-2), we compare the three methods: 'None' (no noise removal), BST-CF and BST-EF. The null hypothesis states that the three methods behave similarly and thus their ranks should be equal. The first step in computing the test is to convert the original

AUC results for each classifier to ranks where the lower ranks, the higher performance in average. The Friedman's average ranks are shown in Table 5.9(a). It is clear that for all classifiers, BST-CF is ranked the first; BST-EF ranked the second; and the last is 'None' with rank 3. The p-values of the Friedman's test are very small for all the classifiers (see Table 5.9(b)), which shows a great significance in the differences found. Thus, the null hypothesis is rejected at a high level of significance.

Table 5.9 : (a) Friedman's average ranks.

method	LR	RF	SVM	AdaBoost
None	3	3	3	3
BST-CF	1	1	1.1	1.3
BST-EF	2	2	1.9	1.7

(b) The Chi-square (with degrees of freedom) and p-value obtained from the Friedman test.

	LR	RF	SVM	AdaBoost
χ^2 (df)	20(2)	20(2)	18.2(2)	15.8(2)
p-value	4.54×10^{-5}	4.54×10^{-5}	0.00011	0.00037

5.3.3 Experiment 3 - Sequential Student Performance Prediction

The results from two experiments are used in the third experiment to predict the response correctness of a student's future trial using only previous session's trials. In this sequential prediction task, trials or instances of dataset are presented sequentially one at time according to student's sessions. The sequential prediction here is a more practical task of prediction where the student's next response needs to be predicted before the student starts the next session. The ML algorithms predict the response correct of the future trial using preceding session's trials as training data. The test data is the trials of one session for all students, and they might have different sizes. In previous experiments, 10-fold cross validation methodology was used and hence each classifier was performed ten runs. Whereas in the current experiment, the classifier executes 15 times and the result is obtained by averaging scores of the 15 runs.

Since weighting and SMOTE-ENN were robust on average among other approaches as we have seen in the results of the previous experiments, we select them to examine our methods BST-CF and BST-EF in experiment 3. Table 5.10 presents the AUC results obtained by different classifiers using SMOTE-ENN and weighting approaches, also including the performance with imbalanced dataset without implementation of any approach 'None'. By comparing with 'None' results, two

approaches have been able to improve the AUC results. The results indicate that the classifiers did achieve better when RT is added to the baseline features. We also observe that RF outperforms the others classifiers (LR, SVM and AdaBoost).

Table 5.10 : The AUC results for prediction based on SMOTE-ENN and weighting.

		SMOTE-ENN		Weighting	
Algorithm	None	Baseline set	Baseline+ RT	Baseline set	Baseline+ RT
RN	0.505	0.501 ± 0.050	0.518 ± 0.073	0.510 ± 0.052	0.503 ± 0.073
LR	**0.557**	0.653 ± 0.042	0.670 ± 0.033	0.651 ± 0.038	0.676 ± 0.039
RF	0.509	**0.675** ± 0.041	**0.684** ± 0.040	**0.667** ± 0.033	**0.689** ± 0.044
SVM	0.541	0.654 ± 0.051	0.671 ± 0.042	0.652 ± 0.049	0.678 ± 0.036
AdaBoost	0.537	0.646 ± 0.039	0.656 ± 0.030	0.655 ± 0.037	0.666 ± 0.039

The best case for each classifier is highlighted in bold.

In this subsection, we examine performances our methods BST-CF and BST-EF to predict the response correct using the best noise set obtained in experiment 2. AUC results for different classifiers after removing noisy instances from DS and DS-2 using BST-CF and BST-EF methods are shown in Table 5.11 and 5.12. The best results are highlighted in bold face. The results indicate that two methods significantly improve AUC scores for all classifiers compared with results obtained in Table 5.10. RF again achieves the best AUC results in all cases. Moreover, the results obtained by two methods on DS-2 are better than results on DS although the size of the dataset has decreased by 22%.

Table 5.11 : The AUC results for sequential prediction based on SMOTE-ENN and using proposed methods.

	with noise eliminated from DS		with noise eliminated from DS-2	
Algorithm	BST-CF	BST-EF	BST-CF	BST-EF
RN	0.510 ± 0.052	0.518 ± 0.045	0.512 ± 0.063	0.529 ± 0.072
LR	0.702 ± 0.044	0.697 ± 0.040	0.730 ± 0.067	0.721 ± 0.048
RF	**0.728** ± 0.046	**0.717** ± 0.038	**0.760** ± 0.049	**0.746** ± 0.035
SVM	0.706 ± 0.038	0.701 ± 0.040	0.733 ± 0.036	0.720 ± 0.041
AdaBoost	0.702 ± 0.034	0.694 ± 0.046	0.718 ± 0.035	0.693 ± 0.053

Table 5.12 : The AUC results for sequential prediction based on Weighting and using proposed methods.

	with noise eliminated from DS		with noise eliminated from DS-2	
Algorithm	BST-CF	BST-EF	BST-CF	BST-EF
RN	0.485 ± 0.060	0.490 ± 0.047	0.514 ± 0.063	0.512 ± 0.072
LR	0.710 ± 0.043	0.705 ± 0.047	0.707 ± 0.067	0.728 ± 0.052
RF	**0.740** ± 0.040	**0.732** ± 0.038	**0.766** ± 0.043	**0.750** ± 0.045
SVM	0.708 ± 0.036	0.705 ± 0.044	0.703 ± 0.070	0.729 ± 0.052
AdaBoost	0.705 ± 0.052	0.690 ± 0.048	0.697 ± 0.061	0.705 ± 0.056

5.4 Important Features on Predicting Student's Performance

In this section, we focused on the important attributes used in predicting student performance. Random forests are used not only for prediction but also to assess the importance of the attributes (Archer and Kimes, 2008). Feature importance evaluates the relative contribution of a feature for predicting the student's response. The importance of a feature $Im(f)$ is computed as the (normalized) total reduction of error brought by that feature. It is also known as the Gini importance. The higher $Im(f)$, the more important the feature. Table 5.13 demonstrates the attributes ranking with their scores of RF's inference task for the dataset. It can be seen that the attribute with the highest rank is RT, and with the lowest rank is category ID. Thus, we can conclude that the attribute RT has a strong impact on RF classifier performance, while category ID has the least impact. When we assess the attributes importance of RF without RT (see Table 5.13), the object ID has a higher rank than others and no change in the ranking of the rest of the attributes' ranking.

Table 5.13 : Importance of attributes using RF model.

	Baseline with RT		Features w/o RT	
Rank	Attributes	Score	Attributes	Score
1	RT	0.187	object_id	0.399
2	student_id	0.326	student_id	0.323
3	object_id	0.299	level	0.213
4	level	0.102	category_id	0.065
5	category_id	0.086		

Figure 5.6 shows the relative importance of all features. One can see clearly that the RT and object's level are the most important factors that determine student's performance. In (Radwan and Cataltepe, 2016) we showed that there is a strong negative correlation between RT and the performance of a student. We also observe that the third ranked feature in figure is student2. In experiment 2, we show that the classifiers perdition is highly affected by instances that belong to student2, because many of them are considered as noise.

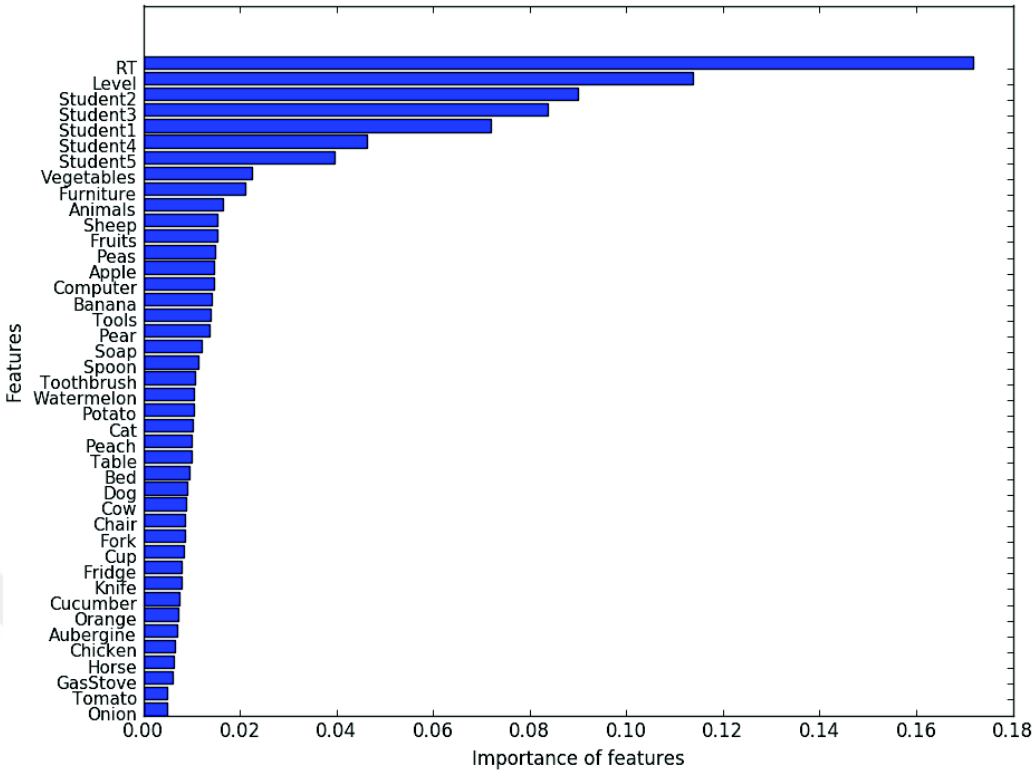

Figure 5.6 : Importance of all features based on RF model.

6. CONCLUSIONS AND FUTURE WORK

The number of children who are diagnosed to have ASD has increased at an alarming rate (Url-3, 2016). In the last decade, a large number of studies have investigated empirical applications of computer-based interventions aimed to be useful for children with ASD. Most of these interventions demonstrated positive effects in improving the social communication ability, emotion recognition and academic learning skills for students with ASD (Smith and Sung, 2014). These studies motivated us to incorporate machine learning models and techniques together with a computer-based application tailored for children with ASD to achieve a further improvement for education of children with ASD.

In the first part of this thesis, we proposed a novel AML framework, called weighted response active machine learning (WRAML), which interactively teaches object recognition to the children with ASD by presenting to the child the most informative teaching set of objects. The study was conducted within an alternating treatments design, with initial baseline and final best treatment phase. For each participant, PL and AML were compared within an alternating treatments phase. Results from this phase demonstrated that the AML was more effective than PL for four out of the five participants. In the best treatment phase for these participants, AML alone was implemented and further increase was observed in the accuracy. Moreover, our AML approach reduces the teaching set size as high as 40% of that required for teaching a child compared with PL.

Future research may include another query selection strategy to choose the informative instances as a teaching set such as Query-by-Committee (Seung et al, 1992) and Expected Error Reduction (Settles, 2010). They have been successfully applied to different classification problems. Moreover, our framework is non-interactive teaching, where the teaching set is computed offline and then presented to the student during teaching session. In future, we plan to develop an adaptive algorithm chooses teaching set online based on the current performance of child during the teaching phase. We think that understanding ML methodologies may provide insights into the

development of new techniques to enhance educational and behavioral therapies for individuals with ASD.

The second part presents an analysis and visualization of the data collected using our web application during the learning sessions. The results of the analysis were interpreted and provided us a better understanding about object categorization for students with ASD. A number of factors influence a student's response including familiarity of categories, difficulty levels, the location of object's image and similarities with the other objects. For the empirical results of the study, we also present justifications to explain these results. The findings help us enhance the teaching process and monitor student's learning progress and then personalize learning for each student. Due to several features for tablet PC, we suggest that the tablet is an effective educational tool to enhance teaching for the students with ASD. Autistic child has much less difficulty interacting with the machine (as opposed to human) and there is no pressure on the child. However, it is difficult to teach a child with autism, and a child gets bored and tired. The lack of emotional interaction between the child and the teacher is a negative aspect. Future research could include using electrodermal activity (EDA), pulse plethesymograph (PPG) and eye gaze to interpret the facial expressions while teaching takes place to determine the internal state of the student. This could be a parameter to the system that offers the next session.

When learning from imbalanced data, classifiers are usually overwhelmed by the majority class, and thus the minority class instances tend to be misclassified. Moreover, the existence of class noise has a greater impact on imbalanced data than on usual cases, and it leads to degradation of classifier performance (Anyfantis et al, 2007; Seiffert et al, 2014). Despite the increasing number of class noise filtering techniques, these have been slightly used in the context of learning from imbalanced data (Sáez et al, 2015). According to our best knowledge, class noise together with class imbalance have not yet been used on an educational domain dataset.

In third part of this thesis, we proposed two empirical methods that deal with class imbalance and noise filtering problems simultaneously. The first method class-Balanced by SMOTE & Thresholding combined with Classification Filter (BST-CF) combines two techniques for class-imbalance with CF using random forest classifier while the second method class-Balanced by SMOTE & Thresholding combined with Ensemble Filter (BST-EF) combines class-imbalance techniques with EF. Our

experimental results on two imbalanced datasets DS & DS-2 show that the performance of a classification algorithms were significantly better with the use of two methods than without them. The two methods can improve the AUC scores, and the second method achieves better TP rate. When we exclude the data of a student with a lot of noisy instances, the prediction results improve. The most robust classifier tested over our methods is the random forest classifier. It performs better on average than logistic regression, SVM and Adaboost with LR. An analysis of feature importance showed that the response time and object's level are the most significant predictors (features) of correctness of a student’s response.

In the future, we aim to carry out the proposed methods on datasets outside the education domain, and examine the effectiveness of them. We will investigate the use of other noise filter approaches such as IPF (Zhu et al, 2003), saturation filter (Gamberger and Lavrac, 1997) or NoiseRank (Sluban et al, 2014) using the same procedures employed in our methods.

REFERENCES

Alberto, P. A., and Troutman, A. C. (2012). *Applied behavior analysis for teachers* (9th ed.). Englewood Cliffs, NJ: Merrill.

Alpaydin, E. (2014). *Introduction to machine learning* (3rd ed). Cambridge, MA: The MIT Press.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders DSM-5. Washington, DC: APA publishing.

Anyfantis, D., Karagiannopoulos, M., Kotsiantis, S., and Pintelas, P. (2007). Robustness of learning techniques in handling class noise in imbalanced datasets. In *Proc. of the IFIP International Federation for Information Processing Comf. AIAI 2007,* 21-28.

Archer, K. J., and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, *52*(4), 2249-2260.

Artoni, S., Buzzi, M. C., Buzzi, M., Ceccarelli, F., Fenili, C., Rapisarda, B., and Tesconi, M. (2012). Designing ABA-based software for low-functioning autistic children. In *Advances in New Technologies, Interactive Interfaces and Communicability Lecture Notes in Computer Science*, 7547, 230-242.

Autism Speaks (2015). Retrieved from http://www.autismspeaks.org/what-autism.

Barlow, D. H., and Hayes, S. C. (1979). Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, *12*(2), 199–210.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A Study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets*, *6*(1), 20-29. http://doi.org/10.1145/1007730.1007735

Berrar, D., and Flach, P. (2011). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Briefings in Bioinformatics*, 13(1), 83-97.

Blagus, R., and Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), 1–16. http://doi.org/10.1186/1471-2105-14-106

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., and Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of Autism and Developmental Disorders*, *45*(5), 1121–1136.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Brodley, C. E., and Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11, 131-167

Bruyer, R., and Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, *51*(1), 5-13.

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., and Zhu, X. (2009). Human active learning. In *Advances in Neural Information Processing Systems* 21 (NIPS 2008), Vancouver, Canada. 241–248.

Chau, V.T.N., and Phung, N.H. (2013). Imbalanced educational data classification: an effective approach with resampling and random forest, In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference.* 135–140.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 341–378. http://doi.org/10.1613/jair.953

Cooper, J. O., Heron, T. E., and Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Corbett, J. (2006). *Health care provision and people with learning disabilities: A guide for health professionals*. Chichester, John Wiley & Sons Ltd.

Cowan, N. (1997). *Attention and memory: An integrated framework*. Oxford Psychology Series (No. 26). New York: Oxford University Press.

Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., and Castiglioni, I. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *Journal of Autism and Developmental Disorders*, 45(7):2146-56. doi: 10.1007/s10803-015-2379-8.

Davy, M. (2005). A review of active learning and co-training in text classification. *Computer Science Technical Report TCD-CS-2005-64, Trinity College Dublin*, 1-39.

Devlin, S., Healy, O., Leader, G., and Hughes, B. M. (2011). Comparison of behavioral intervention and sensory-integration therapy in the treatment of challenging behavior. *Journal of Autism and Developmental Disorders*, *41*(10), 1303–1320.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In: Kittler, J.; Roli, F. (eds.) *Multiple Classifier Systems*. Lecture Notes in Computer Science 1857, 1–15. Springer.

Doenyas, C., Şimdi, E., Özcan, E. Ç., Çataltepe, Z., and Birkan, B. (2014). Autism and tablet computers in Turkey: Teaching picture sequencing skills via a web-based iPad application. *International Journal of Child-Computer Interaction*, *2*(1), 60–71.

Edelson, L., Fine, A., and Tager-Flusberg, H.(2008). Comprehension of nouns and verbs in toddlers with autism: An eye-tracking study. In 7th

International Meeting for Autism Research; London 15–17 May, 2008, Wiley Periodicals, Inc.

Eldevik, S., Ondire, I., Hughes, J. C., Grindle, C. F., Randell, T., and Remington, B. (2013). Effects of computer simulation training on in vivo discrete trial teaching. *Journal of Autism and Developmental Disorders*, *43*(3), 569–578.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *CVPR 2009, IEEE Conference on Computer Vision and Pattern Recognition*, 1778-1785. doi:10.1109/CVPRW.2009.5206772.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. http://doi.org/10.1016/j.patrec.2005.10.010

Fox, J., and Carvalho, M. S. (2012). The RcmdrPlugin. Survival package: Extending the R Commander interface to survival analysis. *Journal of Statistical Software*, 49(7), 1–32.

Galleguillos, C., and Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding* 114(6), 712-722.

Gamberger, D., and Lavrac, N. (1997). Conditions for Occam's Razor applicability and noise elimination. In: Someren, M. V.; Widmer, G. (eds.) *Lecture Notes in Artificial Intelligence:Machine Learning: ECML-97*. 1224, 108–123. Springer-Verlag.

Gamberger, D., Boskovic, R., Lavrac, N., and Groselj, C. (1999). Experiments with noise filtering in a medical domain. In *Proceedings of 16th ICML*, San Francisco, CA, 143-151.

Ganea, P. A., Canfield, C. F., Simons-Ghafari, K., and Chou, T. (2014). Do cavies talk? The effect of anthropomorphic picture books on children's knowledge about animals. *Frontiers in Psychology*, 5: 283. doi:10.3389/fpsyg.2014.00283.

García, S., Fernández, A., Luengo, J., and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10), 2044-2064.

García, V., Sánchez, J. S., Martín-Félez, R., and Mollineda, R. A. (2012). Surrounding neighborhood-based SMOTE for learning from imbalanced data sets. *Progress in Artificial Intelligence*, *1*(4), 347-362.

Gastgeb, H. Z., Dundas, E. M., Minshew, N. J., and Strauss, M. S. (2012). Category formation in autism: can individuals with autism form categories and prototypes of dot patterns?. *Journal of Autism and Developmental Disorders*, *42*(8), 1694–1704.

Gastgeb, H. Z., Strauss, M. S., and Minshew, N. J. (2006). Do individuals with autism process categories differently? The effect of typicality and development. *Child Development*, *77*(6), 1717–1729.

Gibson, B. R., Rogers, T. T., and Zhu, X. (2013), Human semi-supervised learning. *Topics in Cognitive Science*. 5(1):132-72. doi: 10.1111/tops.12010.

Grow, L. L., Kodak, T., and Carr, J. E. (2014). A comparison of methods for teaching receptive labeling to children with autism spectrum disorders: A systematic replication. *Journal of Applied Behavior Analysis*, *47*(3), 600–605.

Guo, X., Yin, Y., Dong, C., Yang, G., and Zhou, G. (2008). On the class imbalance problem. In *Fourth International Conference on Natural Computation IEEE*, 4, 192-201.

Guo, Y. and Schuurmans, D. (2008). Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 20, 593–600.

Han, H., Wang, W. Y., and Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in ımbalanced data sets learning. In *International Conference on Intelligent Computing (ICIC'05)*. Lecture Notes in Computer Science 3644, New York, 878–887.

Harris, H., Israeli, D., Minshew, N., Bonneh, Y., Heeger, D. J., Behrmann, M., and Sagi, D. (2015). Perceptual learning in autism: over-specificity and possible remedies. *Nature Neuroscience*, 18(11), 1574-1576. doi:10.1038/nn.4129.

He, H., and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284.

Hoi, S. C. H., Jin, R., Zhu, J., and Lyu, M. R. (2006). Batch mode active learning and its application to medical image classification. In *Proceedings of the International Conference on Machine Learning (ICML).* ACM, New York, NY, USA, 417-424.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., and Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71(2), 165–179.

Jordan, R. (2013). *Autism with severe learning difficulties* (2nd ed.). London: Souvenir Press.

Kagohara, D. M., Sigafoos, J., Achmadi, D., O'Reilly, M., and Lancioni, G. (2012). Teaching children with autism spectrum disorders to check the spelling of words. *Research in Autism Spectrum Disorders*, 6(1), 304–310.

Khaleefa, O. (2006). Adaptation of the WISC-III in Sudan and Japan: A cross cultural study. *Arapsynet. E. Journal*, No 12, 149–154.

Khoshgoftaar, T. M., Joshi, V., and Seliya, N. (2006). Detecting noisy instances with the ensemble filter: a study in software quality estimation. *International Journal of Software Engineering and Knowledge Engineering*, *16*(1), 53-76. http://doi.org/10.1142/S0218194006002677

Khoshgoftaar, T. M., and Rebours, P. (2007). Improving software quality prediction by noise filtering techniques. *Journal of Computer Science and Technology*, *22*(3), 387-396.

Klinger, L. G., and Dawson, G. (1995). A fresh look at categorization abilities in persons with autism. In E. Schopler & G. Mesibov (Eds.), *Learning and cognition in autism* (pp. 119–136). New York: Plenum Press.

Kosmicki, J. A., Sochat, V., Duda, M., and Wall, D. P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry*, *5*(2), e514.

Lang, R., Ramdoss, S., Sigafoos, J., Green, V. A., van der Meer, L., Tostanoski, A., Lee, A., and O'Reilly, M. F. (2014). Assistive technology for postsecondary students with disabilities. In G. E., Lancioni, N. N. Singh (Eds.), *Assistive Technologies for People with Diverse Abilities* (pp. 53-76). New York: Springer.

Li, H., Zhang, K., and Jiang, T. (2004). Minimum entropy clustering and applications to gene expression analysis. In *Computational Systems Bioinformatics Conference, CSB 2004. Proceedings 2004 IEEE Proceedings*, Stanford, California, USA, 16-19 August, 142–151.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250, 113-141.

Lorah, E.R., Tincani, M., Dodge, J., Gilroy, S., Hickey, A., and Hantula, D. (2013). Evaluating picture exchange and the iPad™ as a speech generating device to teach communication to young children with autism. *Journal of Developmental and Physical Disabilities*, 25, 637–649.

Márquez-Vera, C., Cano, A., Romero, C., and Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, *38*(3), 315-330.

McNaughton, D., and Light, J. (2013). The iPad and mobile technology revolution: Benefits and challenges for individuals who require augmentative and alternative communication. *Augmentative and Alternative Communication*, 29, 107–116.

Mesibov, G. B., and Shea, V. (2010). The TEACCH program in the era of evidence-based practice. *Journal of Autism and Developmental Disorders*, *40*(5), 570–579.

Napierała, K., Stefanowski, J., and Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In: Szczuka, M. et al. (Eds.): *RSCTC 2010, LNAI 6086*, 158–167. Springer-Verlag Berlin Heidelberg.

National Autism Center. (2015). *Findings and conclusions: National standards project, phase 2*. Randolph, MA: (Mills, K.).

O'Malley, P., Lewis, M. E. B., and Donehower, C. (2013). Using tablet computers as instructional tools to increase task completion by students with autism. Online submission, *The Annual Meeting of the American Educational Research Association*, San Francisco, CA, Apr 27-May 1.

Olsson. A. (2009). Literature survey of active machine learning in the context of natural language processing. Technical Report T2009:06, *Swedish Institute of Computer Science*.

Osmanbegović, E., Suljić, M., and Agić, H. (2014). Determining dominant factor for students performance prediction by using data mining classification algorithms. *Transition*, *16*(34), 147-158.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Plaisted, K., O'Riordan, M., and Baron-Cohen, S. (1998). Enhanced discrimination of novel, highly similar stimuli by adults with autism during a perceptual learning task. *Journal of Child Psychology and Psychiatry*, *39*(5), 765–775.

Radwan, A., and Cataltepe, Z. (2016). Using assistive technology to enhance teaching for students with autism spectrum disorders. *International E-Journal of Advances in Education*, *2*(4), 112–121.

Ratcliff, R., and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9(5), 347–356.

Romero, C., and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Romero, C., and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27. http://doi: 10.1002/widm.1075

Sáez, J. A., Luengo, J., Stefanowski, J., and Herrera, F. (2014). Managing borderline and noisy examples in imbalanced classification by combining SMOTE with ensemble filtering. In: Corchado E., Lozano JA., Quintián H., Yin H. (Eds.). *Intelligent Data Engineering and Automated Learning-IDEAL 2014*. Berlin, Heidelberg: Springer, 61-8.

Sáez, J. A., Luengo, J., Stefanowski, J., and Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184-203.

Sarma, B. S., and Ravindran, B. (2007). Intelligent tutoring systems using reinforcement learning to teach autistic students. In Venkatesh, A., Gonzalvez, T., Monk, A., Buckner, B. (eds.) *Home Informatics and Telematics: ICT for The Next Billion.* Proceedings of HOIT 2007. IFIP, vol. 241, 65–78. Springer, Boston.

Satyanarayana, A., and Nuckowski, M. (2016). Data mining using ensemble classifiers for ımproved prediction of student academic performance. In *ASEE Mid-Atlantic Section Spring 2016 Conference*, Washington D.C, April 8-9.

Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., and Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions, Man and Cybernetics, Part A: Systems and Humans*, 40, 185–197.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Folleco, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences*, *259*, 571-595.

Settles, B. (2010). Active learning literature survey. Computer Sciences Technical Report 1648. *University of Wisconsin*–Madison.

Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proc. of the ACM Workshop on Computational Learning Theory*. Pittsburgh, PA, USA, July 27-29.

Sheng, V. S., and Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Massachusetts: AAAI Press. 476-481

Sluban, B., Gamberger, D., and Lavrač, N. (2014). Ensemble-based noise detection: noise ranking and visual performance evaluation. *Data Mining and Knowledge Discovery*, *28*(2), 265-303.

Smith, B. R., Spooner, F., and Wood, C. L. (2013). Using embedded computer-assisted explicit instruction to teach science to students with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 7(3), 433–443.

Smith, V., and Sung, A. (2014). Computer Interventions for ASD. In *Comprehensive guide to Autism*. pp. 2173–2189. Springer. doi: 10.1007/978-1-4614-4788-7_134.

Solomon, M., Smith, A. C., Frank, M. J., Ly, S., and Carter, C. S. (2011). Probabilistic reinforcement learning in adults with autism spectrum disorders. *Autism Research*, 4(2), 109–120.

Sutton, R.S., and Barto, A.G. (1998). *Reinforcement learning: An introduction.* The MIT Press, Cambridge.

Tager-Flusberg, H. (1985). Basic level and superordinate level categorization in autistic, mentally retarded, and normal children. *Journal of Experimental Child Psychology*, 40, 450–469.

Tager-Flusberg, H., and Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: the neglected end of the spectrum. *Autism Research*, 6(6), 468–478.

Thai-Nghe, N., Busche, A., and Schmidt-Thieme, L. (2009), Improving academic performance prediction by dealing with class imbalance, In *Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA 2009)*. Pisa, Italy, IEEE Computer Society, 878 - 883.

Tincani, M. (2004). Comparing the picture exchange communication system and sign language training for children with autism. *Focus on Autism and Other Developmental Disabilities*, *19*(3), 152–163.

Ting, K. M. (1998). Inducing cost-sensitive trees via instance weighting. In: *Proceedings of 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*. Volume 1510 of Lecture Notes in AI, Springer-Verlag, 139–147.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics 6*, 769–772.

Tong, S., and Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 45-66.

Url-1 *<http://mahedee.net/a-mini-spa-with-asp-net-mvc-and-angularjs>*, date retrieved 15.02.2015.

Url-2 *< https://github.com/fmfn/UnbalancedDataset>*, date retrieved 10.02.2016.

Url-3 *<http://www.cdc.gov/ncbddd/autism/data.html>*, date retrieved 31.03.2016.

Vaughan, C. A. (2011). Test review: E. Schopler, M. E. Van Bourgondien, G. J. Wellman, & S. R. Love Childhood Autism Rating Scale (2nd ed.). Los Angeles, CA: Western Psychological Services, 2010. *Journal of Psychoeducational Assessment*, *25*, 489–493.

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., and DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS One*, *7*(8), 1-8.

Walsh, M. B. (2011). The top 10 reasons children with autism deserve ABA. *Behavior Analysis in Practice*, *4*(1), 72–79.

Wang, F. (2014). Learning teaching in teaching: Online reinforcement learning for intelligent tutoring. In James J. *(*Jong Hyuk*)* Park et al. *(*Eds.*)*, *Future Information Technology,* vol. 276, Springer Berlin Heidelberg. 191–196.

Wu, X., and Zhu, X. (2008). Mining with noise knowledge: Error-aware data mining. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, *38*(4), 917–932.

Yang, Q., and Wu, X. (2006). 10 Challenging problems in data mining research. *International Journal of Information Technology; Decision Making (IJITDM)*, *5*(4), 597-604.

Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. (2014). Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, *113*(2), 113–127.

Zhang, C. X., Wang, G.-W., Zhang, J.-S., Guo, G., and Ying, Q.-Y. (2014). IRUSRT: A novel imbalanced learning technique by combining inverse random under sampling and random tree. *Communications in Statistics-Simulation and Computation* 43, 2714-2731.

Zhu, X., and Wu, X. (2004). Class noise vs. attribute Noise: A quantitative study of their impacts. *Artificial Intelligence Review*, 22(3), 177-210.

Zhu, X., Wu, X., and Chen, Q. (2003). Eliminating Class Noise in Large Datasets. In: *Proceedings of the 20th ICML*, 920-927.

Zhu, X., Rogers, T., Qian, R., and Kalish, C. (2007). Humans perform semi-supervised classification too. In Holte, R. and Howe, A. (Eds.), *Twenty-Second AAAI Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, 864-870.

APPENDICES

APPENDIX A: Target Objects.

APPENDIX B: Ethical approval documents.

APPENDIX C: Probability list sample.

APPENDIX A

Table A.1 : List of target categories and objects.

	Category	Object
1	Fruit	apple
		banana
		orange
		pear
		watermelon
		peach
2	Vegetables	tomato
		onion
		cucumber
		potato
		aubergine
		peas
3	Animals	cat
		chicken
		dog
		cow
		sheep
		horse
4	Furniture	computer
		fridge
		gazstove
		bed
		chair
		table
5	Tools	spoon
		fork
		cup
		knife
		soap
		toothbrush

APPENDIX B

İTÜ

İSTANBUL TEKNİK ÜNİVERSİTESİ-SAĞLIK VE MÜHENDİSLİK BİLİMLERİ İNSAN ARAŞTIRMALARI ETİK KURULU (İTÜ-SM.İNAREK)

12 Mayıs 2015

Reference No: BM-28

İstanbul Teknik Üniversitesi Bilgisayar ve Bilişim Fakültesi Bilgisayar Bölümü öğretim üyelerinden Prof. Dr. Zehra Çataltepe yürütücülüğündeki "Otizmli Çocukların Öğretiminde Yapay Öğrenme Yöntemlerinin Kullanımı" adlı proje değerlendirilmiş ve etik açıdan uygun bulunmuştur.

Prof. Dr. Aslıhan Tolun
Boğaziçi Üniversitesi
Moleküler Biyoloji ve Genetik Bölümü

Toplantıya katılmamıştır.
Prof. Dr. İbrahim Akduman
İstanbul Teknik Üniversitesi
Elektronik ve Haberleşme Mühendisliği Bölümü

Prof. Dr. Levent Trabzon
İstanbul Teknik Üniversitesi
Makina Mühendisliği Bölümü

Prof. Dr. Beraat Özçelik
İstanbul Teknik Üniversitesi
Gıda Mühendisliği Bölümü

Doç. Dr. Eda Tahir Turanlı
İstanbul Teknik Üniversitesi
Moleküler Biyoloji ve Genetik Bölümü

Doç. Dr. Gülay Bulut
Bahçeşehir Üniversitesi
Moleküler Biyoloji ve Genetik Bölümü

Doç. Dr. Çiğdem Göksel
İstanbul Teknik Üniversitesi
Geomatik Mühendisliği Bölümü

Uzman Dr. Esra Özaydın
İstanbul Teknik Üniversitesi
Sağlık Kültür Daire Başkanlığı

Pervin Erol, Avukat
İstanbul Teknik Üniversitesi
Hukuk Bürosu

ASLI GİBİDİR

Prof. Dr. Arzu Karabay, **Başkan**
İstanbul Teknik Üniversitesi
Moleküler Biyoloji ve Genetik Bölümü

Mine VURAL
Fakülte Sekreteri

Figure B.1 : Ethical committee approval.

Consent Form

Sponsor group: N/A
Name of study: Otizmli Çocukların Öğretiminde Yapay Öğrenme Yöntemlerinin Kullanımı
Researchers: Prof Dr. Zehra Çataltepe, Akram Radwan
Address: İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi, Maslak, İstanbul, 34469
E-mail address: cataltepe@itu.edu.tr
Phone Number: +90-212-285-3551, +972-599-408042
Researcher name: Akram Radwan
Address: Gaza Strip, Palestine
Phone number: +972-599-408042
E-mail address: aradwan@itu.edu.tr
Researcher name: Dr. Anwar A. Abadshah, Psikolog Dr.
Address: Khan Younis, Gaza Strip, Palestine
Phone number: +972-599-454516
E-mail address: aabadsah@iugaza.edu.ps

Dear Participant,
Prof. Dr. Zehra Çataltepe of the Istanbul Technical University Computer Science Department, her Ph.D. student Akram Radwan, and Dr. Anwar A. Abadshah are organizing a research project under the name "Using Machine Learning to Teach Autistic Children." We invite you to help us gather data for our project. Our aim is to develop software applications to teach various concepts. The data gathered from this research will be used to help understand the learning processes of and possibly cure autistic children.

It is expected that there will be 10 participants diagnosed with autism. Sessions will be conducted at the participants' own schools. The application written will be presented via tablet or personal computer. Participants will attempt to categorize a series of images.

Participation in this project is entirely voluntary. In the event you join, you reserve the right to immediately withdraw without any stated purpose. Neither of the parties involved are obligated to provide monetary compensation to each other.

If you wish to learn more about the research project, you may contact Prof. Dr. Zehra Çataltepe from the Istanbul Technical University Computer Science Department, Science Institute Ph.D. student Akram Radwan, or Dr. Anwar A. Abadshah using the contact information provided above. If you have any questions regarding this project, please ask them before signing this form.

You may contact Prof. Dr. Arzu Karabay (phone number: +90-212 2857257) for more information about your rights regarding this study.

In the event you change your address or phone number, we urge you to inform us.

If you wish to participate in this study, please sign this form.

I, (participant name) ....., have read the above terms and fully understand the scope and purpose of the study, and my voluntary responsibilities regarding the project described. I was able to ask questions regarding the project. I understand that in the event I wish to withdraw from this project, I am free to immediately do so without stating my purpose, without any consequences for myself. I thus choose to participate in this study without coercion and entirely of my own free will. I have read and understood the form, and my questions have been answered. I have received a copy of the form.

Date (day/month/year):...../...../.....

Name-surname of parent/guardian of participant:.....

Signature:.....

Address (include phone/fax number if applicable):……

.....

Name-surname of researcher taking consent:…….

Signature:.....

Name-surname of witness .....

Signature:.....

Figure B.2 : Consent form for parents.

جمعية الحق في الحياة

RIGHT TO LIVE SOCIETY

Dear Sir/Madam,

We confirm that Akram Radwan, who is a PhD Student at Istanbul Technical University, Computer Engineering Department, is allowed to conduct research at our school, under the supervision of Dr. Anwar A. Abadsah.

The research to be conducted is on "Study of Machine Learning Techniques for Teaching Children with Autism". The study will be conducted using tablet PCs on a software written by Akram Radwan. The name of an object will be pronounced on the tablet PC. The student will be asked to select the object among a number of objects shown. The identity of which objects to be shown will be determined based on the students' past performance.

Students' parents/guardians will sign the consent form for the study and will be informed that the participation in the study is voluntary and they can leave the study any time they want. The data gathered during the study will not contain name, date of birth or address of children and can be used for research purposes.

Sincerely,

Name and LastName, Ahmed E. Al Helou
School Principal RLS Executive Manager
Date 4-05-2015
Signature

Right to Live Society
Gaza, Al Sheja'eya.
Tel: +٩٧٠-٨-٢٨٠٧٠١١
Website: www.rightolive.org

T | +970 (8) 280 3888 E | mail@righttolive.org
T | +970 (8) 280 7011 P | P.O. Box: 1021
F | +970 (8) 280 7010

Palestine - Gaza - Shijaia
Al Karama Street
Eastern Road

فلسطين - غـــزة - الشجـاعيـــة
شارع الكرامة (الخط الشرقي) مقابل
مدرســة الأوقـــــاف الشـــرعيـــة

www.rightolive.org

The Winner of | الحائزة على

facebook.com/RLS.Palestine

Figure B.3 : Permission of RTL school.

P.S.A.R

الجمعية الفلسطينية لحالات التوحد والتأهيل

Palestinian Society for Autism & Rehabilitation

Dear Sir/Madam,

We confirm that Akram Radwan, who is a PhD Student at Istanbul Technical University, Computer Engineering Department, is allowed to conduct research at our school, under the supervision of Dr. Anwar A. Abadsah.

The research to be conducted is on "Study of Machine Learning Techniques for Teaching Children with Autism". The study will be conducted using tablet PCs on a software written by Akram Radwan. The name of an object will be pronounced on the tablet PC. The student will be asked to select the object among a number of objects shown. The identity of which objects to be shown will be determined based on the students' past performance.

Students' parents/guardians will sign the consent form for the study and will be informed that the participation in the study is voluntary and they can leave the study any time they want. The data gathered during the study will not contain name, date of birth or address of children and can be used for research purposes.

Sincerely,

Ghada Darwish
Executive Director
16/05/2015
Signature:.....

P.S.A.R

Palestinian Society for Autism & Rehabilitation
Gaza, Remal.
Tel: +970-8- 2860767
Website: www.pal-autism.ps

فلسطين - غزة - الميناء
بالقرب من فندق غزة الدولي
+970 (8) 2860767
pal.autism@hotmail.com

Figure B.4 : Permission of PSAR school.

APPENDIX C

Participant#1 Initial Assessment

Main

Settings

Participants

Dataset

Results

Assessment | Baseline | Confusion Matrix | Passive Learning | Active Learning

Delete

Category	P(Category)	Object	P(Object)
Animals	0.93	Horse	1
		Chicken	1
		Cow	0.8
		Cat	1
		Dog	1
		Sheep	0.8
Fruits	0.83	Orange	1
		Apple	0.8
		Watermelon	1
		Pear	0.6
		Banana	1
		Peach	0.6
Furniture	0.87	Bed	1
		Computer	1
		Table	0.6
		Fridge	0.8
		GasStove	0.8
		Chair	1
Tools	0.97	Fork	0.8
		Toothbrush	1
		Soap	1
		Cup	1
		Spoon	1
		Knife	1
Vegetables	1	Onion	1
		Tomato	1
		Cucumber	1
		Peas	1
		Aubergine	1
		Potato	1

Back to List

Figure C.1 : The initial probability list for participant 1.

CURRICULUM VITAE

Name Surname : Akram Radwan

Place and Date of Birth **:** 16/06/1973

E-Mail **:** akr.radwan@gmail.com

EDUCATION **:**

- **B.Sc.** **:** 1996, Birzeit University, Faculty of Science, Department of Mathematics and Computer Science
- **M.Sc.** **:** 2000, Islamic University of Gaza, Faculty of Science, Department of Mathematics.

PROFESSIONAL EXPERIENCE AND REWARDS:

- 1996 - 2000: Department of Computer and Information, National Center for Information, Gaza, Palestine.
- 2002 – 2010: Lecturer, Department of Information Technology, University College of Applied Science, Gaza, Palestine.
- 2016: Completed Doctorate at İstanbul Technical University

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Radwan, A.**, and Cataltepe, Z., (2016). The Use of Tablet PCs in Teaching Object Recognition to Students with ASD, *Proceedings of 3rd International Conference on Education and Social Sciences INTCESS*, February 8-10, 2016, Istanbul, Turkey, pp. 399-408.
- **Radwan, A.**, and Cataltepe, Z., (2016). Using assistive technology to enhance teaching for students with autism spectrum disorders. *International E-Journal of Advances in Education*, *2*(4), 112–121. DOI: http://dx.doi.org/10.18768/ijaedu.15760
- **Radwan, A. M.**, Birkan, B., Hania, F., and Cataltepe, Z., (2016). Active machine learning framework for teaching object recognition skills to children with autism.

International Journal of Developmental Disabilities. Published online 17 Jun 2016. DOI:10.1080/20473869.2016.1190543

- **Radwan, A.**, and Cataltepe Z., Improving Performance Prediction on Education Data with Noise and Class Imbalance. *Intelligent automation and Soft Computing*. Submitted, July 2016.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- Habil, E., and **Radwan, A.**, (2002). Decomposition of Measures on Difference Posets, *Islamic University Journal*, (10): 75- 90.
- **Radwan, A.**, (2006). Using IPSec in IPv6 Security, *Proceedings of the 4th MultiConference on Computer Science and Information Technology*, Applied Science University, Amman, Jordan, April 5-7, 2006, Vol. 4, pp. 471-474.
- **Radwan, A.**, (2013). Impact of Prerequisite on Prediction of Students' Academic Performance, *Proceedings of The First International Conference on Applied Sciences*, Gaza, Palestine, Sep. 24-26, 2013, pp 105-112.