The Andrew and Erna Viterbi Faculty of ELECTRICAL & COMPUTER ENGINEERING

SIPL SIGNAL AND IMAGE PROCESSING LAB

TECHNION Israel Institute of Technology

דו"ח סיכום פרויקט: א'

זיהוי אוטיזם בעכברים על בסיס ניתוח אותות קוליים

Recognizing Autism in Mice by Analyzing their Squeaks

מבצעים:

איתמר גינסברג	Itamar Ginsberg
אלון שרויאר	Alon Schreuer

מנחה:

דרור לדרמן	Dror Lederman

סמסטר רישום: חורף תשע"ח

תאריך הגשה: ינואר, 2021

בשיתוף עם:

פרופ'ר חווה גולן, אוניברסיטת בן-גוריון

P6222-2-20

תודות

אנו רוצים להודות למנחה דרור לדרמן שעזר וכיוון אותנו לאורך הפרויקט, ובנוסף גם לצוות המעבדה של SIPL ובראשה נמרוד פלג על התמיכה הטכנית בפרויקט. תודה גם לצוותים הקודמים שעבדו על המחקר שבעזרת התוצאות שלהם יכולנו לקדם ולממש את הפרויקט, וכמובן לפרופסור חווה גולן על התמיכה המקצועית והחומרים שלה שהיוו חלק משמעותי מעבודת הפרויקט שלנו.

תוכן עניינים

רשימת איורים

רשימת טבלאות

תקציר

גילוי אוטיזם בקרב ילדים מהווה אתגר משמעותי המעסיק חוקרים רבים. החשיבות בגילוי מוקדם בעלת ערך קרדינלי ביכולת לתת טיפול הולם שיאפשר את שילובם של הלוקים בתסמונת בחברה. עד כה האבחון מתבסס על תצפיות של מאבחנים, כלי זה מוגבל מפני שהוא סובייקטיבי ומתבסס על ניתוח התנהגות שמתפתח בשלבים מעט מאוחרים כמו דיבור ותקשורת, ולכן מקשה על הגילוי המוקדם.

מטרת הפרויקט היא לפתח מודל לגילוי תסמונת הספקטרום האוטיסטי (Autism Spectrum Disorder) בקרב עכברים, המתבסס על האותות הקוליים שלהם. המודל מבוסס על לימוד מכונה – הוא מחשב מאפיינים שחושבו מתוך מאגר נתונים רחב שכולל אלפי הקלטות של עכברים שסווגו לפי הברות. העכברים מחולקים לפי מין, גיל, וכן כאלה שהוזרק להם גן המחקה אוטיזם וכאלה שלא. השלב הראשון של הפרויקט כלל הכנה מוקדמת וביצוע מעבדות אקדמיות בנושא עיבוד אותות דיבור, למידה עמוקה, וקורס בנושא למידת מכונה ולמידה עמוקה של אוניברסיטת סטנפורד. השלב השני של הפרויקט כלל חלק של סקירה ספרותית וקריאה רחבה של מאמרים, לצד העמקה וסנכרון עם העבודה שנעשתה במסגרת המחקר עד כה, מאגר הנתונים הנרחב וקטעי הקוד תוך שחזור תוצאות חיזוי ההברות שהיה הכרחי עבורנו להמשך העבודה. בחלק השלישי היה עלינו לחשוב על מודל ועל מאפיינים משמעותיים שיתנו אפשרות לקבל סיווג איכותי על מנת שהמודל יוכל לתת חיזוי טוב. לכן בחרנו מודל המבוסס על עצי החלטה המתאים למציאת סיווג בינארי, ולכן בחרנו במודל XGBoost שהניב דיוק של 88% ותאם את התכנון שלנו.

Abstract

Diagnosis of autism at an early age is an extensive area of research, as it has a massive impact on the ability to treat and aid those suffering from the syndrome. So far the diagnosis has been based upon professional behavioral observation, a

flawed tool since it is subjective and imprecise, but also due to the fact that it is only effective from a late stage (age 4-5 years).

The goal of this project is to develop a diagnostic-assist tool for classifying mice into two categories: mice with symptoms of ASD (autism spectrum disorder) and mice without such symptoms, based on recordings of their squeaks. The tool is based on the machine-learning methodology. The model we implemented receives features that we extracted from a dataset made up of thousands of mice recordings split into recordings of a single syllable each, and the mice pre-classified by sex, age, etc., and regular mice apart from those injected with a genotype imitating the symptoms of ASD.

We began with a preliminary stage, consisting of academic experiments in the fields of audio signal processing and deep learning at the Technion, and an online review of CS231N – the 'Convolutional Neural Networks for Visual Recognition' course by Stanford University. The second stage included a wide literature review in relevant fields, and synchronization with past projects completed with this research, the dataset used and the code written during them. This stage concluded with a successful regeneration of the results previously achieved. In the third stage we came up with possible models and algorithms for the final classification, as well as choosing prominent data features that will help us successfully classify the mice. After several attempts we agreed upon using XGBoost, an evolution of simple decision-trees algorithms, and after tuning and

optimizing the algorithm, we achieved a final classification accuracy of approximately 88%.

איור 1 - זוג עכברים המתקשרים ביניהם

1. מבוא

1.1. הצגת הפרויקט

עד כה בתחום של גילוי אוטיזם דרך האבחון העיקרית התבססה על תצפיות אנושיות של מאבחנים. דבר זה מעלה שתי נקודות בעייתיות : 1. תצפיות אלה מבוססות על חוות דעת אנושית ולא מהוות אינדיקציה חד משמעית, וזה נתון להרבה גורמים מסיחים והשפעות חיצוניות שעלולות לפגוע באיכות הזיהוי, בנוסף מדובר בזיהוי סובייקטיבי. 2. תצפיות אלה מבוססות על תקשורת ועל תגובות קוגניטיביות לגירויים כמו קריאה בשם ,דיבור וסידור צעצועים, ועל כן בהרבה מקרים היכולת לאבחן רלוונטית רק בגילאים יחסית מאוחרים. מכאן המוטיבציה לפרויקט – כאמור מטרת הפרויקט היא לבנות מודל לאבחון אוטיזם, למעשה בא לתת מענה על הבעיות הקיימות היום ולאפשר אבחון אובייקטיבי. בנוסף המודל יאפשר אבחון בגילאים צעירים של אפילו שנה וזוהי יכולה להיות פריצת דרך משמעותית בכל הנוגע לטיפול באוטיזם, ככל שהגילוי מוקדם יותר ככה האיכות של הטיפול עולה משמעותית. הפרויקט כעת נמצא בשלב שבו העבודה נבדקת על עכברים על מנת לאפשר להסיק מסקנות שניתן יהיה להשליכם על בני אדם.

1.2. מטרת הפרויקט

פרויקט זה הוא חלק ממאמצי מחקר ממושכים שתכליתם לפתח מודל מבוסס למידת מכונה וחילוץ מאפיינים מהנתונים שכבר קיימים לנו ולייצר חזאי שייתן תוצאה בינארית האם קיים אוטיזם או לא, המאפיינים אותם ייצרנו נבחרו לאחר עבודת מחקר וקריאה של מאמרים והם מהווים את הכניסה למודל שבנינו שמוצאו הוא כאמור תוצאת החיזוי.

2. סקר ספרות

2.1. תסמונת הספקטרום האוטיסטי – ASD

תסמונת הספקטרום האוטיסטי (Autism Spectrum Disorder) הוא מונח המתאר מגוון רחב של תסמונות התפתחות נוירולוגיות מורכבות, הגורמות לפגיעה באינטראקציה חברתית ודפוסי התנהגות חריגים. בארצות הברית מאובחן אוטיזם באחד מכל כ-68 ילדים (כ-1.5%)[1], ובישראל אחוז האבחון בקרב ילדים הוא מעט יותר מ-0.5% מהאוכלוסייה. התסמינים הרפואיים העיקריים של התופעה כוללים פגיעה ביכולת החברתית, קושי ביצירת קשרים, תקשורת לקויה, חדגוניות, חוסר יצירת קשר עין, וליקויי דיבור[1].

הגורם המדויק של התסמונת עדיין אינו ידוע, אך גוברת הסברה שהיא נגרמת משילוב של גורמים גנטיים וסביבתיים. אחד מן הגנים המקושרים ביותר לתחום הוא methylenetetrahydrofolate reductase (או בקיצור: MTHFR) שחוסר בו מגדיל את הסיכון לעיכובים התפתחותיים וסימפטומים אוטיסטיים. גורמים סביבתיים שנבדק ביניהם קשר לתופעה הוא גיל מבוגר של הורי הילד, למרות שלא ניתן לדעת האם ההשפעה של גורם זה היא יותר סביבתית או דווקא גנטית (הקשורה לסבירות גבוהה למוטציות).

אבחון בגיל מוקדם מהווה חלק קריטי עבור ילדים הסובלים מן התופעה – הוא מאפשר מחקר רפואי, ייעוץ והתערבות מתאימים במהלך התפתחות הילד, היכולים לכלול טיפולים, תוכניות חינוכיות מתאימות, וטיפול תרופתי מתאים שיכול לעזור בהתמודדות עם סימפטומים. לכל אלו יש השפעות ארוכות-טווח על היכולות היום-יומיות של הילד בתחומי השפה, קשרים בין-אישיים והתנהגויות, דבר המקל משמעותית על איכות חייהם וחיי משפחתם, כמו גם על שילובם בחברה[3].

האתגר המרכזי באבחון ASD הוא המגוון הרחב של הפגיעה בתחומים שונים בילדים הסובלים ממנה. דיוק האבחון גדל ככל שגיל הנבדק עולה, ונכון להיום הגיל הממוצע שבו מאבחנים את

התופעה הוא 4-5, והאבחון נעשה על סמך תצפיות בהתנהגות והתפתחות הילדים על ידי רופאים ופסיכיאטרים מומחים.

בדומה לבני אדם, מחסור בגן MTHFR מעלה את הסיכון להפרעות התפתחותיות וסימפטומים אוטיסטים גם בעכברים, ולכן המודל המסתמך על מדידות בעכברים שימושי וישים גם לבני אדם[3].

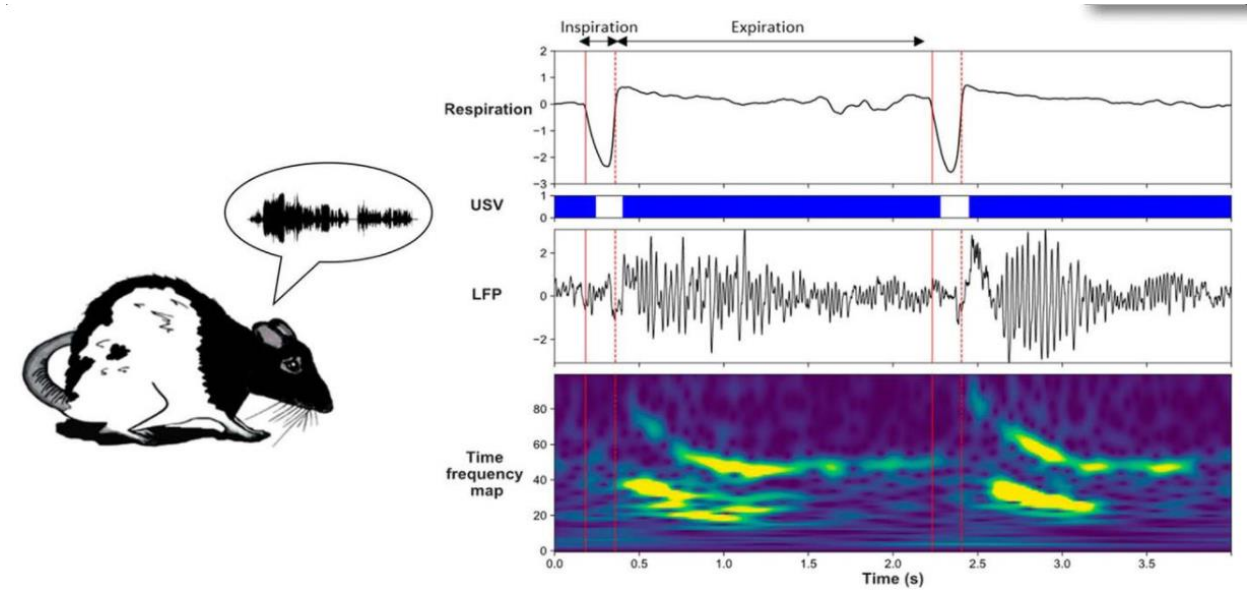
2.2. עיבוד אותות קוליים – Audio Signal Processing

אותות קוליים הם הייצוג החשמלי של קולות ושמע כגון דיבור, מוזיקה וכו'. עיבוד אותות דיבור הוא שם מייצג לכלל הפעולות הממירות את האותות הקוליים לצורה ספרתית/אנלוגית והמבצעות חישובים ושינויים לאות בתחום הזמן ובתחום התדר, כגון סינון, התמרה, אפנון, וכו'[4]. תחום תדרי האותות הנשמעים על ידי האוזן האנושית הוא 20-20,000[Hz]. שימושים מוכרים של עיבוד אותות קוליים כוללים סינון רעשים, דחיסה של מידע, המרה בין אותות אנלוגיים ודיגיטליים, ועוד. בתהליכי חלוקה לרוב נעשה שימוש במאפיינים של אותות קוליים משני התחומים – הזמן והתדר, לדוגמא: Zero Crossing Rate (תדירות חציית האפס של האות בתחום הזמן), Root Mean Squared Energy (שורש הממוצע של ריבוע האנרגיה, המייצג את סך האנרגיה בפרק זמן מסוים), ו-Bandwidth (רוחב הסרט, כלומר פילוג התדרים של האות).

2.3. אותות על-קוליים – USV

אותות המתרחשים בתדרים הנעים בין כ-30-110 קילו הרץ, הנפלטים ע"י בעלי חיים[4]. רוב המחקר שנעשה על אותות הוא על עכברים וחולדות. אותות אלה לא ניתנים לזיהוי ע"י האוזן האנושית, כיוון שהם מחוץ לטווח התדרים הנשמעים על ידי בני אדם, 2-20 קילו הרץ[4]. אותות אלה מהווים כלי תקשורת בין אותם בעלי החיים והם מסווגים לפי תדרים שבהם האותות קיימים.

אותות אלה משתנים לפי מצב רוח רגשי ויכולים להעיד על תפקודים התנהגותיים אצל בעל החיים, וזה הבסיס המחקרי שעליו הפרויקט נשען[4]. אותות אלו ניתן לסיווג לפי סוגי הברות וזהו למעשה בסיס הנתונים של הפרויקט.



איור 2 – המחשה של אותות הדיבור של עכברים בתחום הזמן וספקטרוגרמות במישור הזמן – תדר

2.4. לימוד מכונה – Machine Learning

לימוד מכונה היא כינוי לתחום העוסק בפיתוח אלגוריתמים המיועדים לאפשר למחשב ללמוד מתוך מגוון רחב של דוגמאות, בעל כוח חישובי גדול שנכנס במקומות שבהם התכנות הקלאסי לא מאפשר פתרון.

המטרה המרכזית של למידת מכונה היא טיפול ממוחשב בנתונים מן העולם האמיתי עבור בעיה מסוימת כאשר לא ניתן לכתוב עבורה תוכנת מחשב מפורשת.

אלגוריתמי למידת המכונה מחולקים למספר סוגים נפוצים :

- למידה מונחית (supervised learning) – למידה שמבוססת על תוויות ידועות שמהוות את התוצאה שאליה אנו רוצים להגיע, ולמדוד את איכות המודל שלנו, השגיאה שלנו ושאר

פרמטרים נוספים . למידה זו מכילה בתוכה תהליך של בניית חזאי מתוך מדגם של נתונים שייתן את התוצאה של המודל שבנינו.

למידה זו נפוצה מאוד בבעיות חיזוי.

בנוסף בלמידת מכונה נעשה שימוש בין היתר ברשתות נוירונים מלאכותיות לצורך לימוד וביצוע משימות שונות, בהשראת רשתות נוירונים ביולוגיות של המוח. רשתות אלה בנויות ממודל מתמטי מורכב הממפה באופן לא לינארי בין כניסה ליציאה. הרשת בנויה משכבות, כאשר כל שכבה בנויה ממספר נוירונים אשר מתקשרים ביניהם.

השימוש ברשת נוירונים נפוץ בתחום של למידה מפוקחת (supervised learning). בסוף התהליך מתבצעת פעולת סיווג או סגמנטציה של הדוגמאות וחלוקתן למחלקות שונות. התהליך מתאפשר לאחר אימון משקלי הרשת, המבטאים את הקשרים בין הנוירונים ואת צירופם הכולל לצורך קבלת ההחלטה. אימון הרשת מתבצע ע"י הזנת אוסף של דוגמאות מתויגות לרשת מספר פעמים, כאשר בכל שלב הרשת מקבלת את החלטותיה ומחשבת את שגיאת הסיווג יחסית לתיוג האמיתי. הרשת ממשיכה לעדכן את המשקלים שלה בהתאם לשגיאה, עד להתכנסות לרמת דיוק רצויה לפי סף מסוים. לרוב, שלב אימון הרשת מתבצע על ידי חלוקת אוסף הדוגמאות לשלוש קבוצות: אימון (training), אימות (validation) ובוחן (test). החלוקה מתבצעת אקראית על מנת לשמור על פילוג אחיד של המידע.

סט האימון משמש לצורך אימון הרשת ועדכון המשקלים על מנת למזער את שגיאת הסיווג. סט האימות מהווה דרך לבחינת ביצועי הרשת, וקביעת ההיפר-פרמטרים הרצויים. האימות מתבצע תוך כדי האימון. סט הבוחן נועד להוות בחינה סופית לצורך הערכת ביצועי הרשת בצורה בלתי תלויה בתהליך האימון. ציון הבוחן לרוב יחושב לפי דיוק הרשת (accuracy), כלומר אחוז הדוגמאות שתויגו בצורה נכונה.

אלגוריתמים נפוצים:

- SVM
- Decision Trees
- K-NN
- שימוש במשערכים כגון – MAP,MLE

- למידה בלתי מונחית (unsupervised learning) – למידה שבה התוויות אינן ידועות, מהווה שם כולל למגוון של בעיות שבהינתן מדגם נרצה ללמוד את תכונות המדגם. בניגוד ללמידה מונחית המדגם מכיל את אוסף הדגימות בלבד ללא התוויות.

  דוגמאות לבעיות מסוג זה :

  - חלוקה לאשכולות
  - מציאת ייצוגים טובים לדגימות
  - דחיסה
  - למידת פילוג הדגימות
  - אנומליה(חריגה)

- למידה מחיזוקים – אלגוריתם המקבל משוב חלקי על הביצועים שלו ומנסה לבנות סוכן שיסיק מסקנות על אילו מההחלטות שלו הביאו אותו להצלחה או כישלון.

3. הפתרון

3.1. מאגר הנתונים

מאגר הנתונים שלנו מכיל כ-4880 הקלטות מתויגות של כ-66 עכברים (כתוצאה מתהליך של הרחבת מאגר הנתונים המתויגים במהלך העבודה), כל אחת מהן מכילה הברה בודדת שהופרדה וסווגה על ידי פרוייקטים קודמים. ההקלטות מכילות מידע על שם העכבר, מין, תיוג גנוטיפ האם, תיוג גנוטיפ העכבר, זמן ההתחלה וזמן סיום ההקלטה, תדר התחלה וסיום, מספר ההקלטה, הזמנים שבין הברה להברה הקודמת (ISI), משך זמן ההברה, וסוג ההברה (שמה ומספרה).

B	C	D	E	F	G	H	I	J	K	L
Day	Minute	Path	Mother	Mother Genotyp	Name	Sex	Offspring Genot	Genogroup	Recording Numb	Start Point (s)
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1		
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	0.088414
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	0.27279
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	0.457
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	0.63737
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	0.83246
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	1.0379
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000001	1.2375
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	0.11691
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	0.32287
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	0.50782
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	0.69659
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	0.90275
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	1.103
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	1.3088
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	1.4887
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000002	1.675
4	1	rs\ella9\Desktop	26598N-1741O	WT	17470O	F	WT	1	T0000003	0.30677

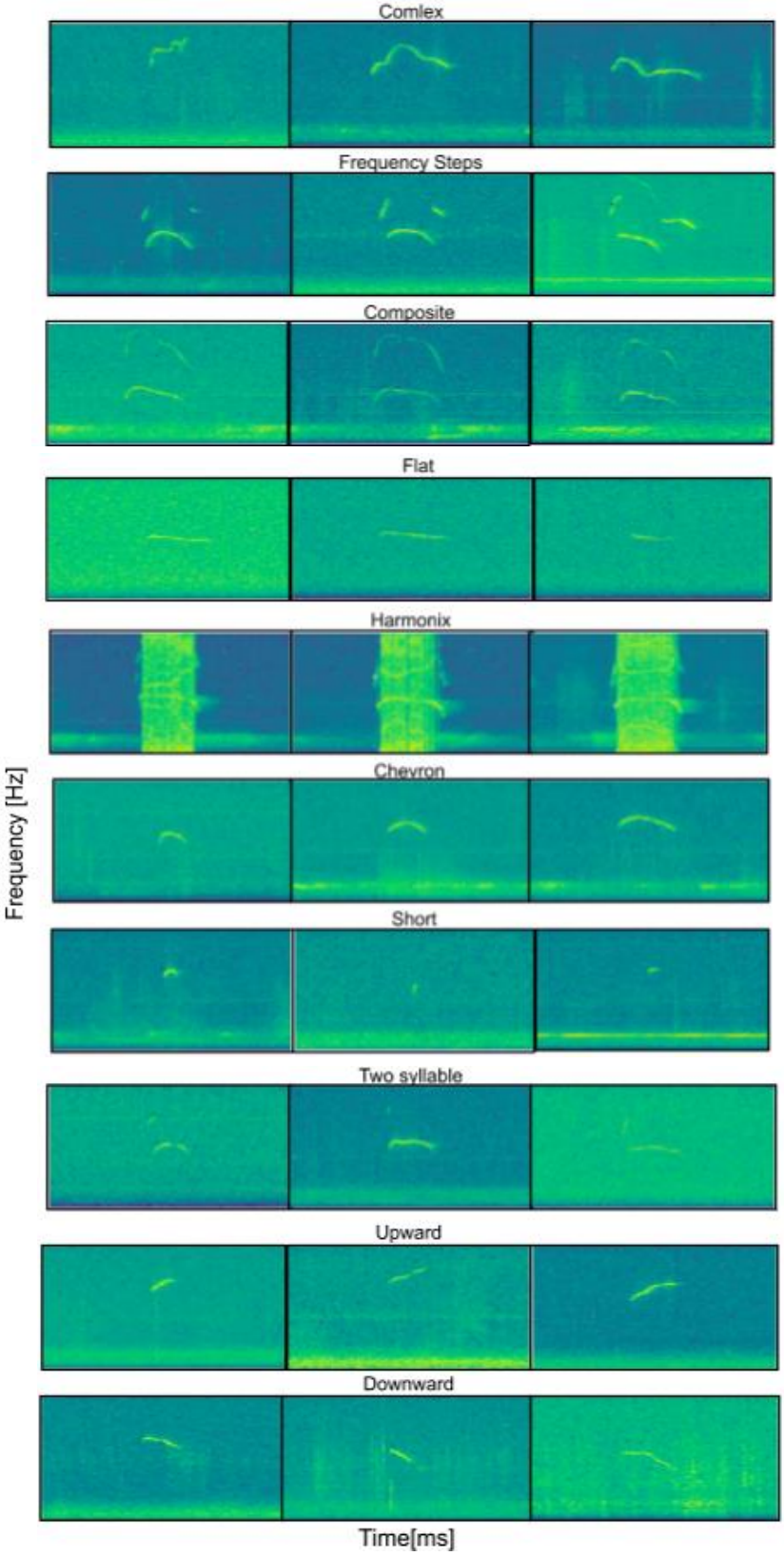
טבלה 1 – מאפיינים חלקיים של מאגר הנתונים

סיווג ההברות מחולק ל-10 קבוצות, וזוהי החלוקה שלהן במאגר:

שם ההברה	מספר הדגימות במאגר
Short	260
Chevron	279
Complex	703
Flat	34
Downward	15
Upward	34
Two Syllable	841
Frequency Steps	1484
Composite	1122
Harmonics	108

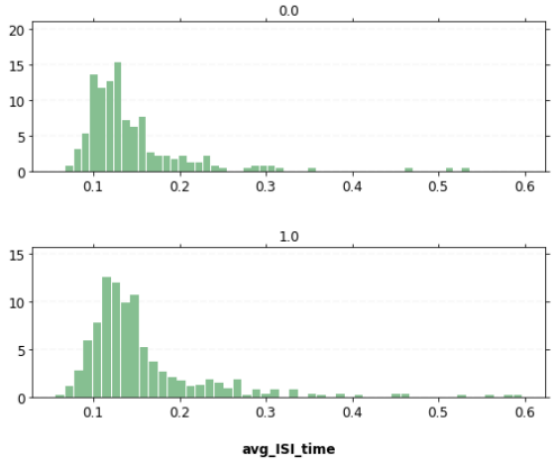
טבלה 2 – רשימת ההברות ומספר הדגימות של
כל אחת במאגר הנתונים

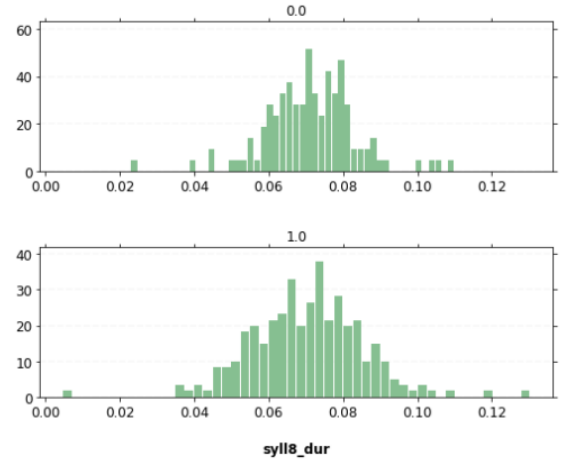
מתוך 4880 ההקלטות, כ1415 שייכות לעכברים המציגים סימפטומים של אוטיזם, וכ-3465 הנותרות שייכות לעכברים הבריאים. ניתן לראות שגם עבור הקטגוריה הזו וגם עבור פילוג ההברות, מאגר המידע אינו מאוזן.

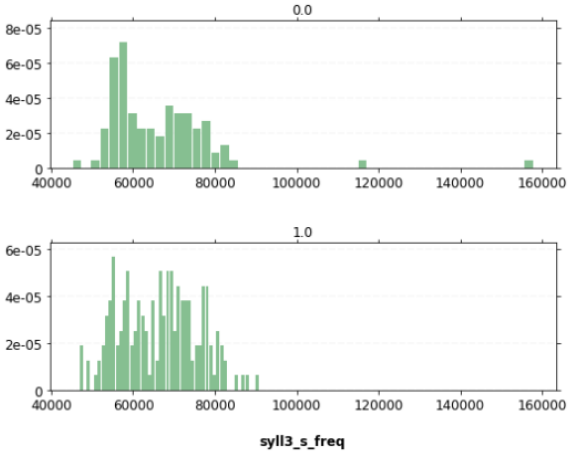


איור 3 – ההברות המסווגות מיוצגות
בספקטרוגרמות

ניתן לחלץ מהנתונים מגוון רחב של מאפיינים, למשל: התפלגות ההברות בקטע דיבור, משך זמן הדיבור, תדרים שדיבור בהם עלול לסמן על אוטיזם, רוחב סרט ועוד רבים. מתוך מאפיינים אלה ניתן למצוא הבדלים בין מחלקות העכברים המאפשרים סיווג והפרדה ביניהם.







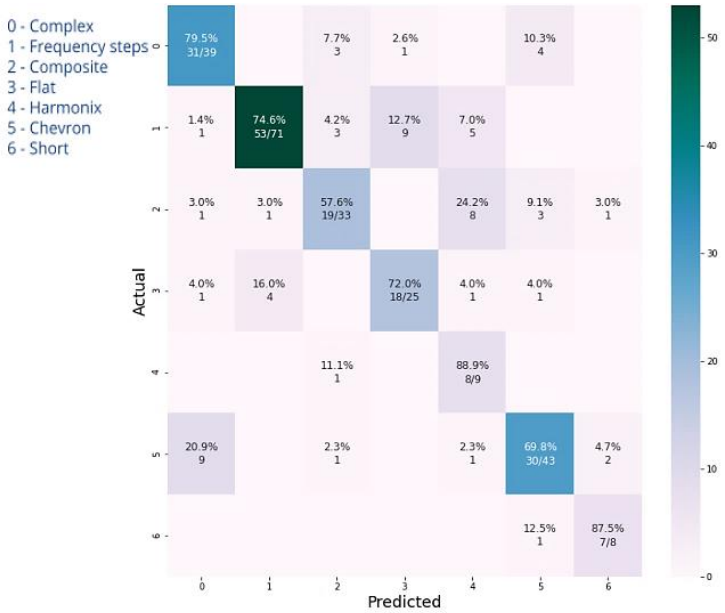
איורים 4,5,6 – דוגמאות לפילוגי המאפיינים בין שתי המחלקות השונות. הכותרת של כל תרשים מציינת את מספר המחלקה שמוצג פילוג המאפיין שלה (0 = עכברים בעלי מאפייני אוטיזם, 1 = עכברים בריאים).

איור 4 (שמאלי עליון) מתאר את היסטוגרמת משך הברה מס' 8 (Frequency Steps).
איור 5 (ימני עליון) מתאר את היסטוגרמת ממוצע הזמן בין הברות בהקלטה.
איור 6 (שמאלי תחתון) מתאר את פילוג תדרי ההתחלה של הברה מס' 3 (Complex).

3.2. עבודות מקדימות

הפרויקט שלנו מתבסס על פרויקטים קודמים שבוצעו על-ידי סטודנטים מאוניברסיטת בן-גוריון והמכון הטכנולוגי חולון במסגרת המחקר. בפרויקטים אלו פותח מודל לסגמנטציה וסיווג של ההברות השונות. מודל זה מתבסס בעיקרו על אלגוריתמים של למידה עמוקה לזיהוי סוג ההברות על ידי ייצוג האותות האודיטוריים באמצעות ספקטרוגרמות וסיווגן באמצעות רשת קונבולוציה עמוקה. הצוותים הצליחו לסווג בדרך זו את ההברות בדיוק של כ-85%. באיור מספר 7 ניתן לראות

מטריצת מבוכה של תוצאות הסיווג, המתארת את אחוז ומספר הדגימות לפי סיווג המודל אל מול הקטגוריה האמיתית שלהן.



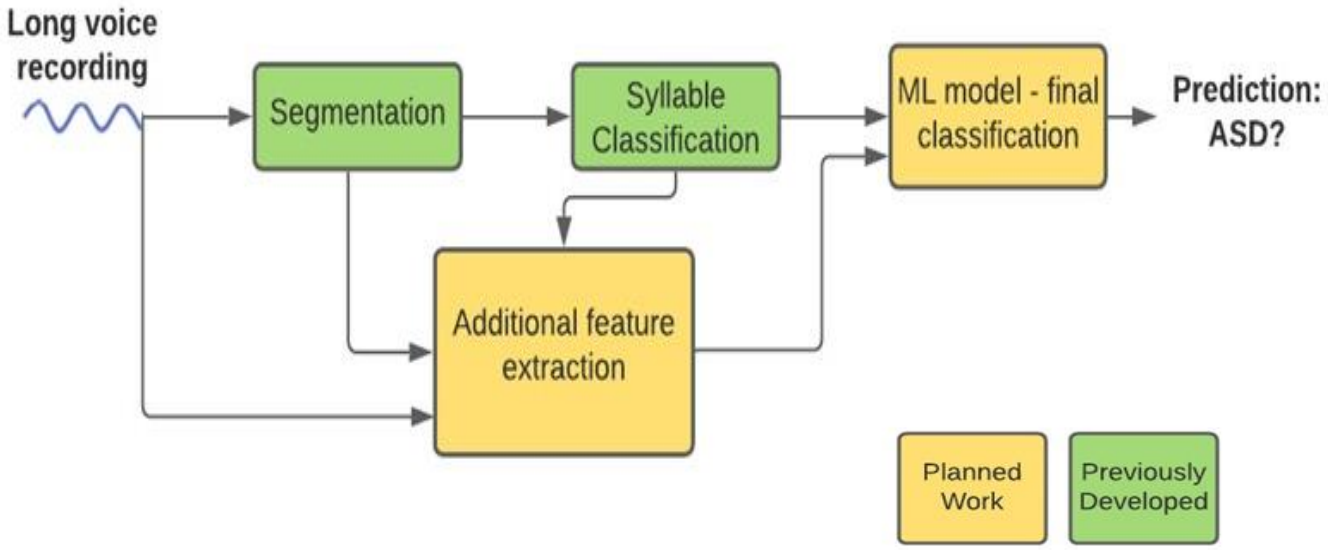
טבלה 3 – מטריצת מבוכה המתארת את ביצועי מודל סיווג ההברות

3.3. תהליך בחירת הפתרון

המודל השלם

על מנת להשלים את משימת חיזוי הגנוטיפ של כל עכבר, בחנו מספר אפשרויות פעולה. אחת העיקריות שבהן הייתה לעבוד עם מודלים כמו LSTM (ארכיטקטורת רשת נוירונים נשנית - RNN) העובדים על רצף שלם של נתונים כגון ההקלטות הישירות עצמן, ללא צורך בעיבוד שנעשה קודם לכן על ידי הצוותים הקודמים (end-to-end). בחרנו שלא להשתמש בגישה זו, על מנת לשמור על רציפות בין פרויקטים כחלק מן המחקר, ועל מנת להשתמש במודלים פשוטים יותר המאפשרים לנו יותר שליטה על כשלים אפשריים בדרך. בנוסף, מאגר הנתונים הקיים אינו בסדר

גודל המתאים לשימוש בשיטות של למידה עמוקה – מספר הדגימות קטן מאוד. איור מספר 9 מציג סכמת בלוקים של המודל השלם המוצע.



איור 7 - דיאגרמת בלוקים המתארת את המודולים של הפרויקט, אלה שנעשו בשלבים קודמים ואלה שנעשו בשלב הנוכחי

בחרנו להתמקד בחילוץ מאפיינים נוספים מתוך מאגר הנתונים ובמודל הסיווג הסופי, מתוך הרצון לפתח מודל אבחון ראשוני שיאפשר לקבל תוצאות סיווג לאוטיזם, שניתן להעריך את הביצועים שלו, ורק לאחר מכן לשפרו בעתיד.

מכיוון שידענו שהמידע שאנו מקבלים בסיום התהליך הקיים מתקבל בצורת טבלה, מימשנו אלגוריתמים המאפשרים חילוץ המאפיינים בהתאם, וחיפשנו מודלים המתאימים לעבודה עם מידע טבלאי. אחד המודלים שנחשב למתקדם ביותר בתחום זה ומציג ביצועים גבוהים באתגרי רשת בנושאי למידת מכונה הוא XGBoost.

3.4. מודל הסיווג הנבחר – XGBoost

הפתרון שבחרנו לממש הוא מודל מבוסס למידת מכונה שיקבל מאפיינים מתוך מאגר הנתונים ויידע לתת חיזוי בינארי באשר לכל עכבר, האם הוא מאובחן כבעל אוטיזם או לא. בנקודה שבא

התחלנו את הפרויקט היה ברשותנו את מאגר הנתונים מחולק לפי סגמנטים שבהם מתרחש אות הדיבור, ותיוג של כל ההברות. כחלק מהפתרון בחרנו באלגוריתם XGBOOST (Extreme Gradient Boosting), אלגוריתם למידת מכונה המבוסס על מספר עצי החלטה ועושה שימוש במסגרת להגברת שיפוע בבעיות הנוגעות לחיזוי נתונים כמו תמונות, אותות דיבור או נתונים טקסטואליים[6]. שימוש במודל זה מאפשר לנו להשתמש בעבודות קודמות וקח לייצר הדרגתיות בעבודה, ויכולת ניטור תקלות גבוהה יותר, וכפועל יוצא גם תורם לנו להבנה טובה יותר של פעולתו.

איור 8 - המחשה למודל XGBOOST המורכב
מאוסף של עצים

3.5. שימוש במודל

3.5.1. מאפייני האותות הקוליים

לאחר עבודת מחקר וסקירה ספרותית, החלטנו לממש את המאפיינים שיהיו הכי אינפורמטיביים ויניבו את התוצאות הטובות ביותר ביעילות[5]. מאפיינים אלה הם:

- התפלגות הברות בהקלטה – מיוצגת על ידי 10 מספרים המהווים את אחוז הקריאות בהברה זו מתוך ההקלטה
- ממוצע תדרי התחלה של הברה – מיוצג על ידי 10 מספרים, אחד עבור כל הברה
- ממוצע תדרי הסיום של הברה – מיוצג על ידי 10 מספרים, אחד עבור כל הברה
- משך זמן הברה – מיוצג על ידי 10 מספרים, אחד עבור כל הברה
- תורשה – גנוטיפ אם העכבר – מיוצג על ידי מספר יחיד, 0 – HT, 1 – WT
- מין - מיוצג על ידי מספר יחיד, 0 - נקבה, 1 – זכר
- זמן הממוצע בין הברות (ISI) – מיוצג על ידי מספר יחיד

בסך הכל מדובר ב-7 מאפיינים המיוצגים על ידי 43 מספרים.

3.5.2. היפר-פרמטרים של מודל הסיווג

לאחר שימוש בכלים כמו חיפוש רשתי (Grid Search) על טווח ערכים רחב ותוך ניסוי וטעייה, בחרנו את ערכי ההיפר-פרמטרים הבאים:

שם הפרמטר	תיאור	ערך מספרי נבחר
Train-test split	חלוקה לסט אימון ובוחן. מתאר את אחוז הדגימות המשמשות כסט הבוחן	0.2
N_estimators	מספר עצי ההחלטה שהמודל מייצר בתהליך האימון	50
Learning rate	קצב הלימוד של המודל בין עצי החלטה עוקבים	0.1
Max depth	העומק המקסימלי לכל עץ החלטה שהמודל מייצר	5
Colsample by tree	החלק היחסי של המאפיינים המשמש בחישוב ויצירה של כל עץ	1
Reg lambda	משקול איבר רגולריזצייה מסוג L2	1.5
Reg alpha	משקול איבר רגולריזצייה מסוג L1	0.05

Scale pos weight	משקול מתעדף לסיווג מוצלח של דגימה מתוך המחלקה המכילה פחות דגימות במאגר הנתונים. פרמטר זה מאפשר להתמודד עם מאגר נתונים לא מאוזן, המכיל יחס לא שווה בין דגימות בין המחלקות	0.8

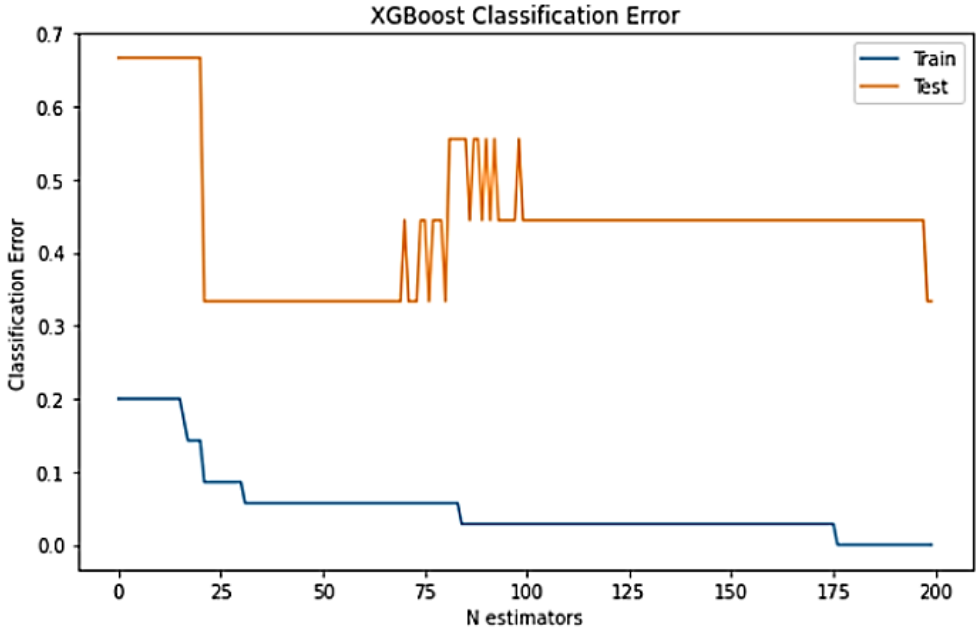
טבלה 4 - ערכי ההיפר-פרמטרים
הנבחרים למודל

4. תוצאות ומסקנות

4.1. תוצאות

4.1.1. אבחון בלתי-תלוי בנבדק

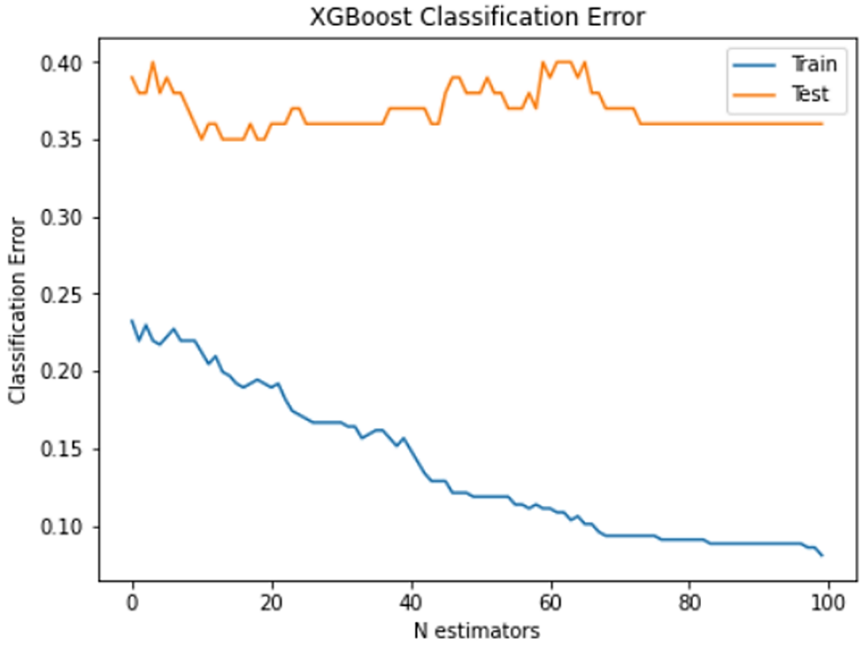
בשלב הראשון, נעשה ניסוי אבחון "בלתי-תלוי בנבדק" (Subject-independent). בניסוי זה נעשה קיבוץ של כל הנתונים על פי המספר הסידורי של כל עכבר – כלל הקלטות ההברות של כל עכבר קובצו למעין "הקלטה אחת" שעליה חושבו כלל המאפיינים (למעט פילוג ההברות, כלל המאפיינים חושבו כממוצע על פני כלל ההקלטות, ובניסיון אחר כחציון). לפיכך קיבלנו סט דגימות בגודל של 44 (כמספר העכברים). בשני הניסיונות האלה אחוז הדיוק על סט האימון היה 97-100% ועל סט הבוחן 66%.



איור 9 - שרטוט שגיאת הסיווג לאורך תהליך האימון, עבור הניסיונות הראשונים

4.1.2. אבחון תלוי-נבדק

נוכח כמות הדגימות הקטנה המתקבלת בכל קבוצה בניסוי הראשון, בניסוי השני בחנו גישה שונה, באופן תלוי בנבדק (Subject-dependent) – התייחסנו לכל מספר הברות של עכבר מסוים, שהוקלטו כחלק מאותו מקבץ הקלטות, כהקלטה יחידה של עכבר (גם כאן תוך לקיחת ממוצע על פני כלל המאפיינים, למעט פילוג ההברות). מטרת הגישה הייתה להגדיל את מאגר המידע, ואכן



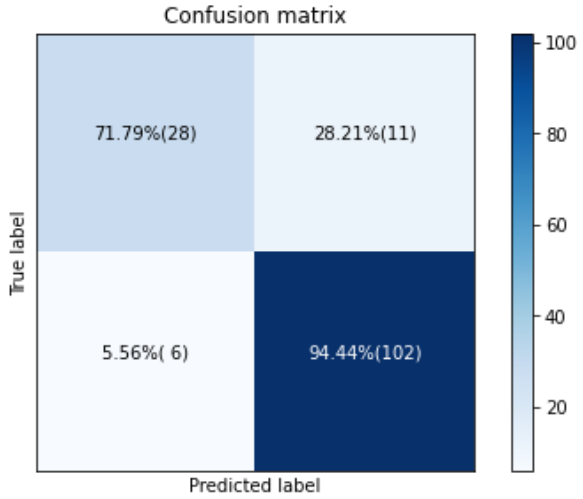
איור 10 – שרטוט שגיאת הסיווג
לאורך תהליך האימון, עבור
הגישה השנייה

כתוצאה מכך הסט הכולל שלנו גדל ל-731 דגימות. ללא שינוי משמעותי של הפרמטרים, אחוזי הדיוק נשארו דומים (64-66%). בנוסף ניתן לראות התאמת-יתר (Over-fitting) של המודל, כלומר פגיעה ביכולת ההכללה על סמך התאמת-יתר של המודל לסט האימון. התופעה באה לידי ביטוי בפער בין שגיאת האימון לבין שגיאת הבוחן, כפי שניתן לראות בתרשים.

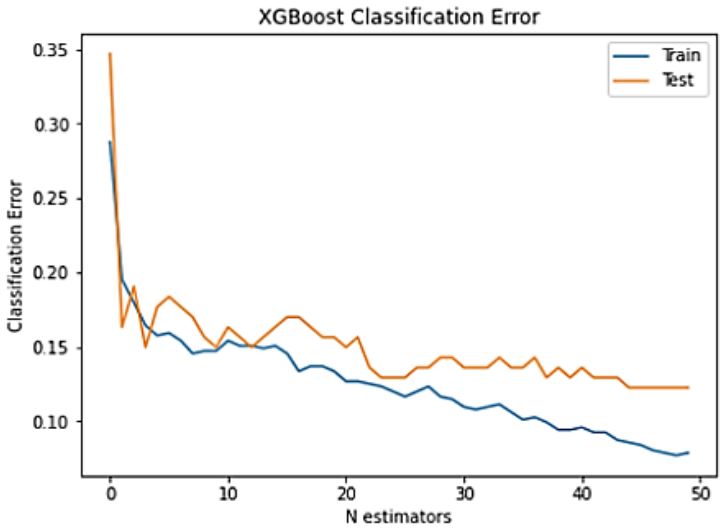
4.1.3. שיפורים נוספים

לבסוף, הוספנו את שלושת המאפיינים האחרונים המייצגים את גנוטיפ האם, מין העכבר המוקלט והזמן הממוצע בין ההברות שלו. לאחר אופטימיזציה של כלל הפרמטרים הנדרשים, שימוש

ברגולריזציה מסוג L1 ו-L2, וכיוונון הפרמטרים המפצים על חוסר האיזון בין מספר הדגימות לכל קטגוריה, הגענו לתוצאות חיזוי של 93% על סט האימון ו-88% על סט הבוחן.



איור 11 – מטריצת המבוכה
של מודל הסיווג הסופי



איור 12 – שרטוט שגיאת הסיווג
לאורך תהליך האימון

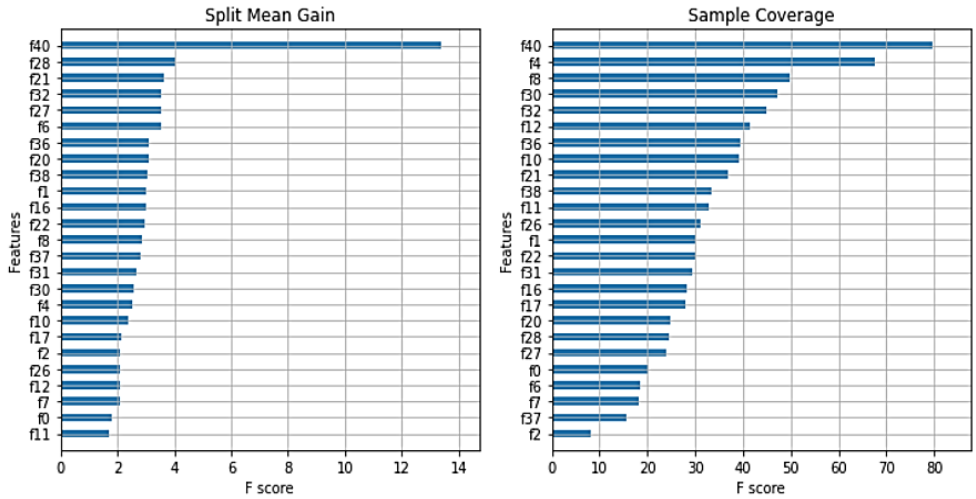
הגרף השמאלי מראה את שגיאת הסיווג על סט האימון וסט הבוחן לאורך תהליך האימון, כלומר ככל שיותר עצי החלטה מחושבים. מימין ב"מטריצת המבוכה" (Confusion Matrix) ניתן לראות את אחוזי הסיווג הנכון והלא נכון לפי קטגוריה – מרבית הדגימות (130 מתוך 147 בסט הבוחן) סווגו לקטגוריה המתאימה, בעוד שה-17 הנותרות סווגו לא נכון (11 דגימות שייכות לקבוצת העכברים האוטיסטים וסווגו על ידי המודל כבריאים, ו-6 דגימות שייכות לקבוצת העכברים הבריאים וסווג על ידי המודל כאוטיסטים).

4.2. מסקנות

4.2.1. חשיבות המאפיינים

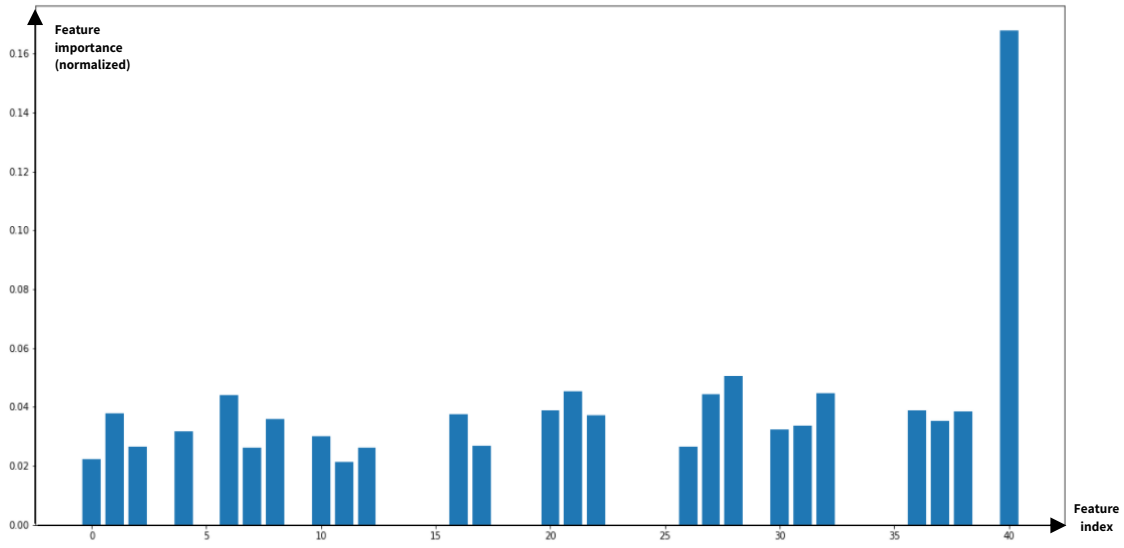
אחד מן המאפיינים המשמעותיים ביותר לחיזוי הגנוטיפ של העכבר המוקלט הוא גנוטיפ האם.

לשם השוואה עם התוצאות שקיבלנו, אימנו מודל זהה לחיזוי גנוטיפ העכבר על סמך המאפיין הזה בלבד: התוצאות שהתקבלו היו 64% אחוזי דיוק על סט האימון ו-73% על סט הבוחן. התוצאה פחות טובה לעומת תוצאות החיזוי שלנו עם שימוש במאפיינים המחולצים מניתוח ההקלטות, ומכאן הערך המוסף שלהם. המחשה נוספת לעובדה זו ניתן לספק על ידי זיהוי המאפיינים המשמעותיים ביותר בתהליך הסיווג מתוך המודל.



איור 13 – תוצאות ניתוח חשיבות המאפיינים
במודל, המסודרים בסדר יורד

ניתן לראות בבירור שהמאפיין המשמעותי ביותר הוא גנוטיפ האם (f40), ולאחריו בסדר חשיבות יורד המאפיינים הקוליים שחולצו מן ההקלטות, הבולטים שבהם הם: ממוצע תדר הסיום של הברה 7 (two-syllable), תדר ההתחלה של הברה 1 (short), ממוצע משך הברה 7, ותדר ההתחלה של הברה 7 (ניתן לראות שמתוך עשרת המאפיינים השימושיים ביותר, ארבעה מכילים מידע הקשור להברה מס' 7) בנוסף ניתן לראות שישנם מאפיינים שלא הייתה להם חשיבות כלל – בין אם כי אין בהם מידע שימושי לסיווג, או כי לא נאספו מספיק דגימות המכילות אותם (כגון הברה מס' 6 – Upward, או 5 – Downward).



איור 14 – המחשה נוספת לסדר
חשיבות המאפיינים בתהליך הסיווג

4.2.2. המלצות לעבודת המשך

כחלק מן הרעיונות שהועלו לאורך תהליך העבודה וכתוצאה מן המסקנות שלנו, אנו מציעים מספר דרכים להעמיק את הניתוח:

- מתוך המסקנות על החשיבות היחסית בין מאפיינים, ניתן למקד את המודל עוד על ידי צמצומו באמצעות שיטות המגדירות ערך סף ומנכות מאפיינים בעלי ערך חשיבות נמוך ממנו מן המודל, כך שמתקבל מודל קומפקטי. המלצתנו לסף ראשוני מתוך התוצאות – 0.03.
- חילוץ מאפיינים נוספים, בעיקר בתחום התדר, כגון רוחב פס (Bandwidth) של כל הברה.
- שימוש בשיטות של למידה עמוקה ורשתות נוירונים – לשם כך נדרש להגדיל משמעותית את מאגר המידע - ככל שנמשך המחקר, ניתן להשתמש בתוצרי פרויקטים קודמים המאפשרים לבודד ולסווג הקלטות חדשות שמתווספות עם הזמן.

5. סיכום

כאמור מטרת הפרויקט היא בניית מודל לימוד מכונה שייתן חיזוי האם העכבר לוקה בתסמונת אוטיזם או לא, שלטובתו חילצנו מאפיינים מתוך מאגר נתונים שיזין את המודל. הפרויקט חולק למספר שלבי עבודה אותם תיארנו בתחילת הדוח. הפתרון שבחרנו הוא בניית מודל XBGOOST המבוסס על עצי החלטה שאפשר לנו עבודה מבוקרת בשלבים שעושה שימוש בעבודות קודמות, תוך חישוב מאפיינים מתוך מאגר הנתונים שנבחרו לאחר סקירה ספרותית וקריאת מאמרים בנושא[5]. לאחר אימון המודל והזנתו במאפיינים שבחרנו הצלחנו להגיע לתוצאות של 88% זיהוי.

אנו מרגישים כי הפרויקט תרם לנו רבות בכך שלימד אותנו מהי עבודת מחקר ומהו תהליך של לימוד מכונה וניתוח נתונים. רכשנו כלים חשובים בתחום לימוד מכונה וידע לא מבוטל בתחום של ניתוח אותות דיבור, וכן העמקנו את הידע שלנו בכל הקשור לתסמונת האוטיזם, והאבחון שלה. נהנו מאוד במהלך הפרויקט ,הרגשנו שאנחנו חלק מתהליך ארוך וגדול של עבודה מעמיקה והתחברנו מאוד למטרה שעומדת מאחורי כל זה. אנו שמחים על התוצאות שקיבלנו על הדרך שעשינו ועל ההשקעה הרבה, ומקווים שפרויקט זה יהווה נדבך חשוב בהמשך קידום המחקר ופיתוח הכלים לאבחון אוטיזם בגיל מוקדם.

6. מקורות

[1] A. Masi, M. DeMayo, N. Glozier, and A. Guastella, “An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options,” vol. 33, no. 2, pp. 183–193, 2017, doi: 10.1007/s12264-017-0100-y.

[2] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the ABIDE dataset,” vol. 17, p. 16, 2018, doi: 10.1016/j.nicl.2017.08.017.

[3] N. Sadigurschi and H. M. Golan, “Maternal and offspring methylenetetrahydrofolate-reductase genotypesinteract in a mouse model to induce autism spectrum disorder-like behavior,” vol. 18, no. 1, p. n/a, 2019, doi:10.1111/gbb.12547.

[4] A. H. O. Fonseca, G. M. Santana, B. O. Gabriela,M., S. Bampi, and M. O. Dietrich, “Analysis of ultrasonic vocalizations from mice using computer vision and machine learning,” vol. 10, 2021, doi: 10.7554/eLife.59161.

[5] H. M. Golan, “USV_MTHFR_CPF_MS”, *Frontiers in Neuroscience* vol.15, p. 11/22/2021, doi : 769670

[6] Tianqi Chen ,Carlos Guestrin,” XGBoost: A Scalable Tree Boosting System”, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, doi: 10.1145/2939672.2939785