Speaker Notes — Analysis of USVs for Autism Detection

מצגת הגנה · Chen Aharon & Aviel Bitton · דף עזר למציג (לא מוקרן)

Slide 1 — **Chen · 0:30**

פתיחה והצגת הפרויקט

- חן פותחת: לברך את הבוחן/ים והמנחה, להציג את שנינו בשם.
- משפט אחד על הנושא: can ultrasonic calls of mouse pups flag an ASD genotype?
- לציין שהמנחה הוא ד"ר דרור לדרמן, ושהנתונים של ד"ר חוה גולן מאונ' בן-גוריון.
- **מעבר:** "אבל לפני המדע — למה בכלל נכנסנו לזה." ומעבירה לאביאל.

Slide 2 — **Aviel · 1:30**

המוטיבציה האישית

- אביאל מספר בגוף ראשון, קצר ומכבד — בלי שם, בלי פרטים מזהים.
- מישהי קרובה מאוד אליי אובחנה על הרצף בגיל שלוש.
- ברגע שהתאימו לה את הסביבה לאיך שהיא חווה את העולם — היה שיפור ניכר בהתנהגות ובתקשורת.
- המסר: ההתאמה המוקדמת היא קריטית, אבל האבחון היום איטי וסובייקטיבי.
- מכאן נולדה השאלה — האם אפשר לזהות מוקדם יותר, ובאופן objective?
- **מעבר:** "וזו בדיוק הבעיה המדעית." → שקופית הבעיה.

Slide 3 — **Aviel · 1:30**

הבעיה — אבחון התנהגותי

- אבחון אוטיזם היום הוא התנהגותי — דורש קלינאי מיומן וזמן רב.
- הכי קשה דווקא בינקות המוקדמת — וזה בדיוק החלון שבו התערבות הכי משמעותית.
- לכן יש ערך עצום בbiomarker אובייקטיבי, כמותי ומוקדם.
- הסטטיסטיקה (1 מ-36) רק כדי למקם את גודל הבעיה — לא להתעכב.
- **מעבר:** "איפה מוצאים אות כזה? מסתבר — אצל עכברים." → שקופית ה-USV.

Slide 4 **Aviel · 2:00**

למה עכברים / למה USV

- גורי עכברים שמופרדים מהאם פולטים קריאות על-קוליות בטווח 35–125 kHz — לא נשמעות לאוזן אנושית.
- זו התנהגות תקשורתית מוקדמת, טבעית, ולא-פולשנית — מודל מצוין לפנוטיפ תקשורתי.
- המודל הגנטי: Mthfr. הגור הוא HET (מודל-ASD, המחלקה החיובית) או WT (ביקורת).
- להצביע על הספקטרוגרמה: כל תיבה כחולה = הברה אחת שהמערכת זיהתה בהקלטה אמיתית.
- **מעבר:** "ולקריאות האלה יש צורה — שהגנוטיפ מעוות." → הטקסונומיה.

Slide 5 **Aviel → Chen · 2:00**

לקריאות יש צורה

- יש עשרה סוגי הברות — מ-upward/downward sweeps ועד harmonic stacks. רשת CNN מסווגת כל הברה.
- ההבדל בין WT ל-HET הוא בעיקר בתדרי הקצה ובמשך ההברה — כפי ש-Shekel 2021 דיווחו.
- שני כיוונים שנבדוק: (1) אגרגציה אקוסטית לכל הקלטה, (2) הסדר הזמני של הקריאות (Gal 2023).
- שני הכיוונים האלה יהפכו לשני מסלולי המודלים — tabular ו-sequence.
- **מעבר לחן:** "אבל לפני שנמדל — צריך להבין מאיפה התחלנו, וזה סיפור בפני עצמו." חן ממשיכה.

Slide 6 **Chen · 1:30**

שאלות המחקר

- ארבע שאלות. שלוש מדעיות ואחת הנדסית — וזו לא פחות חשובה.
- RQ1: האם המודלים החדשים (TabPFN, רשתות רצף) מנצחים את ה-XGBoost שירשנו.
- RQ2: אילו מאפיינים אקוסטיים מניעים את הסיווג.
- RQ3: איך משטר ההערכה (מוכר מול לא-מוכר, pooled מול per-strain) משפיע — זה הלב.
- ההנדסית: לקחנו pipeline שבור והפכנו אותו למערכת רפרודוסבילית עם מקור-אמת יחיד.
- **מעבר:** "בואו נתחיל מאיפה שהתחלנו — וזה לא היה מקום טוב." → מה שירשנו.

Slide 7 Chen · 2:00

מה שירשנו — פרוטוטייפ שבור

- נקודת ההתחלה: קוד מונוליטי שירשנו — מודל אחד בלבד, subject-dependent XGBoost.
- נתיבים קשיחים, אפס תיעוד, לא רץ מ-clean checkout. ישן כשנתיים וחצי.
- המסר החשוב: **כל** מה שמימין — אנחנו בנינו. תיבה כתומה אחת ירשנו, כל השאר חדש.
- זה ממצב את התרומה: גם מדעית וגם הנדסית.
- **מעבר:** "איך הגענו מזה לזה? זה היה מסע." → ציר הזמן.

Slide 8 Chen · 1:45

ציר הזמן — המסע

- שוחזר מהיסטוריית ה-git — כרונולוגיה בלבד, בלי מספרי ביצועים.
- מהפרוטוטייפ הבודד (2023), דרך ההשתלטות שלנו (סוף 2025) ובניית ה-pipeline מחדש.
- נקודה קריטית: כשהוספנו subject-independent split גילינו דליפת מידע — וגם את שגיאת הגנוטיפ.
- ואז נוספו המודלים: TabPFN, רשתות הרצף, ניתוח per-strain, וכיול הסף.
- **מעבר:** "שתי התחנות ב-2026-04 הן סיפור בפני עצמו — מקור אמת יחיד." → מקור האמת.

Slide 9 Chen · 2:00

מקור אמת יחיד + תיקון הגנוטיפ

- כל האימון נשען על קובץ קנוני אחד עם manifest מגורסר — במקום גיליונות מפוזרים ולא מתועדים.
- כל שלב סינון וכל ספירת שורות מתועדים — הכל ניתן לשחזור ולביקורת.
- תוך כדי האצירה גילינו שגיאת תיוג: גורים קיבלו את גנוטיפ האם במקום את שלהם.
- תיקנו **14 עכברים / 2,495 שורות** מ-HET ל-WT — הצלבת HET×WT נותנת חצי גורים WT, אז זה תיקון גנטי הכרחי.
- זה בדיוק מה שיסביר למה ה-baseline שלנו שונה מ-0.829 — נגיע לזה בתוצאות.
- **מעבר:** "ואת האצירה הזו אפשרנו עם כלי שבנינו לחוקרת." → האפליקציה.

Slide 10 Chen · 2:00

אפליקציית ה-Segmentation

- בנינו אפליקציית דסקטופ עבור ד"ר חוה גולן — כדי שתוכל לאצור ולסקור את ה-USV בלי שורת פקודה.
- היא עוטפת בדיוק את אותו pipeline — segmentation + CNN typing — ומייצרת את הקובץ הקנוני שכל המודלים צורכים.
- להראות את עץ השנים (2015–2024), בחירת התיקיות, וה-progress — זה כלי אמיתי, לא מוקאפ.
- זה גם הכלי שדרכו התגלתה שגיאת הגנוטיפ — האצירה קרתה במקביל למידול.
- תוצר אמיתי: installer ל-Windows, קוד פתוח, ופורסם ב-Zenodo עם DOI.
- **מעבר לאביאל:** "עכשיו כשהנתונים נקיים ומסודרים — מה עושים איתם." אביאל ממשיך למסע הנתונים.

Slide 11 Aviel · 2:15

מסע הנתונים — הצינור המלא

- זה עמוד השדרה של כל העבודה — כדאי להישאר עליו רגע.
- איסוף: מערכת Avisoft ב-250 kHz, בפרוטוקול בידוד בימים P4–P12.
- עיבוד: גלאי segmentation על מסנן gammatone, ואז רשת CNN שמסווגת כל הברה לאחד מ-10 סוגים.
- התוצאה: 125,576 הברות — המצע המשותף לשני המסלולים.
- מסלול tabular: אגרגציה לכל הקלטה, 48 פיצ'רים, 12,323 שורות → מודלי עצים ו-TabPFN.
- מסלול sequence: שומר את הסדר, 408 sessions → שלוש רשתות נוירונים.
- ושניהם נכנסים לאותו פרוטוקול הערכה — וזה מה שנפרק עוד רגע.
- **מעבר:** "אבל קודם — כמה מילים על הנתונים עצמם, כי המבנה שלהם קובע הכל." → הנתונים.

Slide 12 Aviel · 2:00

הנתונים — לא מאוזנים ולונגיטודינליים

- המספרים: 126 גורים, 35 אמהות, 5 שנות הקלטה, ~125 אלף הברות.
- חוסר איזון של ~3:1 לטובת WT — לכן accuracy לבד מטעה (ניחוש הרוב נותן 0.75). נסתכל על recall/precision/AUC.
- המבנה הלונגיטודינלי: כל גור תורם הרבה הקלטות מתואמות — חציון 828 הברות לגור.
- המשמעות הקריטית: פיצול נאיבי מכניס את אותו עכבר גם ל-train וגם ל-test → ניפוח מלאכותי.
- לכן חייבים לפצל **לפי עכבר** — וזה מוביל ישר לשקופית ההערכה.
- **מעבר:** "אבל לפני ההערכה — אילו מודלים בכלל הרצנו, ולמה TabPFN." → המודלים.

Slide 13 **Aviel · 2:00**

המודלים — ולמה TabPFN

- שני עולמות ב-tabular: XGBoost שירשנו — עצים עם gradient boosting, המודל-בסיס.
- ו-TabPFN — החידוש המתודולוגי: transformer שאומן מראש על מיליוני טבלאות סינתטיות.
- בזמן שימוש הוא לא מתאמן בכלל — מקבל את ה-train וה-test יחד ב-forward pass אחד, in-context learning. אפס כיול על הנתונים שלנו.
- להצביע על הסכמה: משמאל האימון האיטרטיבי של XGBoost, מימין ה-inference הקפוא של TabPFN.
- שישה מודלים בסך הכל: אחד ירשנו (כתום), חמישה חדשים (ירוק).
- **מעבר:** "ועכשיו — השאלה שקובעת איך קוראים את כל התוצאות." → ההערכה.

Slide 14 **Aviel · 2:30**

ההערכה — שתי שאלות

- זו השקופית הכי חשובה מבחינה מתודולוגית — כדאי להאט כאן.
- Subject-dependent (כחול): פיצול אקראי, אותו עכבר יכול להופיע ב-train וב-test. עונה: כמה טוב על הקלטה חדשה של עכבר **מוכר**. זה ה-best-case, וזה מה שהעבודות הקודמות דיווחו.
- Subject-independent (כתום): פיצול **לפי עכבר**, אף עכבר לא משותף. עונה: כמה טוב על עכבר **לגמרי חדש**. זה מבחן ההכללה האמיתי.
- המסר: ה-independent תמיד נמוך יותר — וזה **לא** כישלון, זו התשובה לשאלה קשה יותר.
- הבחירה לדווח את שניהם זה בדיוק מה שמבדיל את העבודה הזו — שקיפות מלאה.
- **מעבר:** "עם המסגרת הזו — הנה מה שמצאנו." → תוצאות.

Slide 15 **Aviel · 2:45**

תוצאה מרכזית — TabPFN מנצח

- כל dumbbell מחבר את הציון על עכבר **מוכר** (כחול) לציון על עכבר **לא-מוכר** (כתום).
- TabPFN הכי טוב בשני המשטרים — עונה חיובית על RQ1: 0.781 מוכר, 0.729 לא-מוכר (wF1 0.743, AUC 0.783).
- נקודה יפה: הדיוק שלו על עכבר **לא-מוכר** (0.729) שווה בערך לדיוק של המודל הישן על עכבר **מוכר** (0.733).
- והכי חשוב קלינית: tuned-XGBoost שומר HT recall 0.869 על עכברים לא-מוכרים — כמעט לא מפספס גור אוטיסט. לכלי סקר, זה הדבר.
- "הכי טוב" תלוי במטרה: TabPFN לדיוק כולל, tuned-XGBoost לרגישות למיעוט.
- **מעבר:** "אבל רגע — היה מספר קודם, 0.829. מה קרה לו?" → פירוק ה-0.829.

Slide 16 Aviel · 2:45

פירוק ה-0.829

- זו שקופית לתרגל הכי הרבה — כאן שאלות צצות.
- בספרות הקודמת של הקו הזה דווח 0.829 (within-subject, לפני התיקון).
- שלב 1 — תיקון תיוג הגנוטיפ מוריד את ה-baseline של המודל הירושה ל-0.733. זה **לא** רגרסיה — המספר הישן היה מנופח בגלל תוויות שגויות. 0.733 הוא הבסיס הגנטית-נכון.
- שלב 2 — המודל החדש (TabPFN) חוזר ועובר: 0.781 within-subject.
- ועל עכברים לא-מוכרים: XGBoost 0.693, TabPFN 0.729 — הפער הצפוי בין השאלות, לא פגם.
- המשפט לזכור: "המספר הכן נמוך יותר — ואנחנו עדיין מנצחים אותו."
- **מעבר:** "איך אנחנו יודעים שהדירוג אמיתי ולא תלוי-סף?" → ROC.

Slide 17 Aviel · 1:45

ROC — דירוג בלתי-תלוי-סף

- ROC-AUC = ההסתברות שהמודל מדרג HET אקראי מעל WT אקראי — בלי תלות בסף החלטה.
- TabPFN מוביל: 0.908 על מוכר, 0.783 על לא-מוכר.
- ה-AUC יורד מ-within ל-across אצל כל מודל — הפער הזה **הוא** מחיר ההכללה.
- הערה כנה: הכיוונון שיפר recall אבל לא את הדירוג — ה-AUC של tuned אפילו יורד מעט מתחת ל-XGBoost הרגיל על לא-מוכר.
- אם ישאלו על ה-kink בעקומת ה-independent: זה נובע ממעט ציוני-סף ייחודיים ב-test קטן של עכברים לא-מוכרים.
- **מעבר:** "אבל יש קיר — וזה החלק הכי מלמד." → precision wall.

Slide 18 Aviel · 2:15

קיר ה-precision

- זו התוצאה השלילית הכי מלמדת — ואנחנו מציגים אותה בגאווה, זו כנות מדעית.
- בגרף: לכל הברה, ההתפלגויות של WT ו-HET כמעט חופפות לחלוטין — בתדרים ובמשך.
- התוצאה: HT precision תקוע סביב 0.50 אצל **כל** מודל ובכל משטר — כמחצית מהחיוביים הם התרעות שווא.
- כי זה שורד כל אסטימטור — זה מאפיין של ה**פיצ'רים**, לא של המודל.
- הפתרון: פיצ'רים עשירים יותר לכל קריאה (קונטור תדר, רוחב פס, עומק אפנון) — לא מודל גדול יותר. זה גם ה-future work המרכזי.
- **מעבר:** "ניסינו משהו שאפתני כדי לשבור את זה — רשתות רצף." → הניסיון החדשני.

Slide 19 Aviel · 2:00

הניסיון החדשני — רשתות רצף

- זה חלק שאני גאה בו גם אם התוצאה שלילית — ניסינו משהו שאפתני.
- ההשערה: Gal 2023 הראו שהדינמיקה הזמנית של הקריאות נושאת מידע על הפנוטיפ.
- אז דיברנו כל session כרצף מסודר של הברות (עד 256 צעדים, 14 מימדים) — לא רק ממוצעים.
- שלוש ארכיטקטורות: BiLSTM, 1D-CNN, Transformer — רקורנטיות, קונבולוציה, ו-attention.
- וחיפוש שיטתי: 8 מנופים לחוסר-איזון × 3 מודלים × 2 משטרים — בערך 36 ריצות.
- התקווה: שה"תחביר" הזמני ישבור את קיר ה-0.50.
- **מעבר:** "אבל כשבדקנו את זה כמו שצריך..." → מה קרה.

Slide 20 Aviel · 2:00

רשתות רצף — התוצאה

- הריצה הבודדת הכי טובה נתנה balanced accuracy 0.704 (independent) — נראה מבטיח.
- אבל תחת 5-fold grouped CV זה נפל ל-0.563 ± 0.063 — כלומר ה-0.704 היה **fold ממוזל**, לא אות אמיתי.
- הסיבה: סקאלת נתונים — רק ~19 sessions של HET ב-test, מול רשתות של עשרות-אלפי פרמטרים.
- המסגור הכן: זו **לא** הפרכה של Gal 2023 — אלא "לא נתמך בסקאלת הנתונים הזו".
- עם יותר חיות, או אגרגציה לכל עכבר, ייתכן שהסדר כן יעזור — future work.
- **מעבר:** "ויש עוד אזהרה אחת, אולי הכי חשובה." → strain confound.

Slide 21 Aviel · 2:15

ה-strain confound

- אנחנו **בעצמנו** מסמנים את זה — זו האזהרה הכי חשובה, ובוחן יעריך שאנחנו מקדימים אותו.
- strain1 (2022–24, רקע מעורב) מגיע ל-~0.90 על עכברים לא-מוכרים — אבל זה cohort חדש ונפרד יותר.
- הבעיה: strain, שנה, ושכיחות ה-HET כולם משתנים יחד עם הגנוטיפ — אז המודל אולי קורא **שייכות ל-cohort** ולא את הפנוטיפ.
- על strain2 (רק 47 עכברים) מחלקת המיעוט **קורסת** — וה-accuracy (0.65) מסתיר את זה.
- לכן המספר שמייצג אותנו הכי טוב הוא ה-**pooled** על עכברים לא-מוכרים — הוא ממצע את ה-confound.
- המבחן המכריע — לאמן על strain אחד ולבחון על השני — הוא future work מוצהר.
- **מעבר לחן:** "אז מה בעצם אפשר לומר על USV ואוטיזם?" חן לסיכום.

Slide 22 Chen · 2:00

מסקנות — תשובות לשאלות

- לחזור על כל שאלת מחקר עם תשובה קצרה — לסגור מעגל מול שקופית 6.
- RQ1: כן — TabPFN מנצח בשני המשטרים; tuned-XGBoost שומר recall גבוה למיעוט.
- RQ2: תדרי-קצה ומשך נושאים את האות, אבל ה-precision של המיעוט תקוע ~0.50 ברמת הפיצ'רים.
- RQ3: מוכר ≠ לא-מוכר; הסדר לא עזר בסקאלה הזו; per-strain מבולבל עם cohort.
- הנדסי: pipeline רפרודוסבילי, מקור אמת יחיד, אפליקציית אצירה, ו-baseline מתוקן וניתן לביקורת.
- הכותרת הכנה: מודל within חזק (~0.78–0.80) והכללה מעל-מקרה לעכברים לא-מוכרים (0.729, AUC 0.78) — והכל ניתן לשחזור.
- **מעבר:** "ולאן ממשיכים מכאן." → future work.

Slide 23 Chen · 1:30

Future work

- מסודר לפי ערך צפוי יורד — כל פריט קשור למגבלה שהראינו.
- (1) הכי חשוב: פיצ'רים עשירים לכל קריאה — כדי לשבור את קיר ה-precision.
- (2) מבחן cross-cohort אמיתי — לאמן על strain אחד, לבחון על השני — כדי להפריד פנוטיפ מ-cohort.
- (3) החלטה לכל עכבר — הצבעה אחת לחיה, הרזולוציה הקלינית.
- (4) יותר נתונים לרשתות הרצף. (5) ואימות ה-front end.
- **מעבר:** "תודה — ואנחנו פתוחים לשאלות." שניהם קמים. → סיום.

Slide 24 Chen & Aviel · 0:30

סיום ו-Q&A

- משפט סיכום אחד: בנינו מערכת רפרודוסבילית ונתנו תמונה מכוילת של מה ש-USV יכולים ולא יכולים.
- להודות למנחה ד"ר לדרמן ולד"ר חוה גולן.
- שניכם עומדים יחד לשאלות. לנשום. הנספח מוכן מאחורי השקופית הזו.

Slide A1 Q&A · —

טבלת התוצאות המלאה

- לשלוף כשנשאלים "מה המספרים המדויקים". חיובי = TabPFN. HET מודגש.

Slide A2 — · Q&A

חשיבות פיצ'רים + confound של גנוטיפ האם

- שקופית הגילוי הכן — לשלוף אם שואלים "מה הפיצ'ר הכי חשוב" או "אולי המודל קורא רק את האם".
- גנוטיפ האם שולט ב-gain כי הוא חוזה **מעצם ההצלבה**: אם WT → כל הגורים WT; רק HET יכולה להוליד HET.
- כלומר המודל ה-tabular מקבל stratification "בחינם" ממטא-דאטה של יוחסין — לא מהקולות.
- כשמוציאים מטא-דאטה, תדרי-הקצה והמשך מובילים — עקבי עם הספר ועם Shekel 2021.
- לכן אנחנו נשענים על independent / AUC / precision-wall — הם משקפים את האות האקוסטי האמיתי. ablation בלי מטא-דאטה הוא future work.

Slide A3 — · Q&A

Confusion matrices

- לשלוף על "איפה הטעויות". השורות מנורמלות ל-recall; הטעות השלטת = WT שסווג HET (false positive), עקבי עם קיר ה-0.50.

Slide A4 — · Q&A

מטריצת הניסויים

- לשלוף על "כמה שיטתי היה". כל מודל × cohort × split. תאים = test accuracy; לזכור ש-baseline של הרוב ~0.75.

Slide A5 — · Q&A

המודל הטוב ביותר לכל תרחיש

- לשלוף על "מי הכי טוב איפה". שים לב ל-strain1-independent 0.903 — מרשים אך מבולבל עם cohort (ראה שקופית 21).

Slide A6 — · Q&A

כיול סף החלטה

- לשלוף על "אפשר לשפר את ה-precision עם סף אחר". כן — within-subject עד ~0.62; across-subject עדיף (recall ~0.86, precision ~0.46) target-recall. אבל זה לא מזיז AUC.