סמינר בנושא מיוחד במדעי המחשב

אלגוריתם היער העמוק

מחבר: שאול אברג'יל

מנחה: ד"ר מיה הרמן

ספטמבר 2017

תוכן עניינים

[**מבוא** 3](#_Toc493871732)

[**עצי החלטה** 3](#_Toc493871733)

[**יער אקראי** 5](#_Toc493871734)

[Bootstrap 5](#_Toc493871735)

[Bagging 5](#_Toc493871736)

[אלגוריתם היער האקראי 6](#_Toc493871737)

[יתרונות היער האקראי 7](#_Toc493871738)

[גרסאות שונות 7](#_Toc493871739)

[**ארכיטקטורת עמוקות** 9](#_Toc493871740)

[מבוא ללמידה עמוקה 9](#_Toc493871741)

[שימושים נפוצים 12](#_Toc493871742)

[מגבלות 12](#_Toc493871743)

[**יער עמוק** 13](#_Toc493871744)

[אפיון מבנה היער העמוק 13](#_Toc493871745)

[תיאור האלגוריתם 13](#_Toc493871746)

[תהליך האימון 15](#_Toc493871747)

[Multi-Grained Scanning 16](#_Toc493871748)

[מבחני ביצוע 19](#_Toc493871749)

[הרחבות אפשריות 20](#_Toc493871750)

[**סיכום** 21](#_Toc493871751)

[**מקורות** 22](#_Toc493871752)

# מבוא

למידת מכונה הוא תחום שבו תוכניות מחשב לומדות לפתור מטלות, מבלי להיות מתוכנתות באופן מפורש כיצד לעשות זאת. אלגוריתמים של למידת מכונה מיושמים במגוון של עולמות תוכן כמו עיבוד שפה, עיבוד תמונה, חיזוי ערך מניות וכדומה. הסמינר הנוכחי נועד להציג אלגוריתם למידת מכונה חדשני הנקרא יער עמוק.

יער עמוק הינו אגד של עצי החלטה [1, 2], המשובצים יחדיו במבנה רב-שכבתי שמאפשר יכולות למידה מורכבות. הטופולוגיה של יער עמוק מתבססת על ארכיטקטורות של אלגוריתמים מתחום הלמידה העמוקה [8, 10], שיתרונן היחסי הוא בלמידת ייצוגיות והפשטה גבוהה של הנתונים [9]. ייחודיותו של היער העמוק מתבטאת ביכולתו להתמודד עם מגבלות שעדיין אין להן מענה מלא בעולם הלמידה העמוקה: הצורך בנפח רב של נתונים מתויגים[[1]](#footnote-1), והמיומנות הנדרשת לשם כוונון הפרמטרים האלגוריתמיים.

השוואה בין ביצועיו של היער העמוק ואלגוריתמי למידה עמוקה, הן מבחינת זמני ריצה והן מבחינת תוצאות בפועל, תומכות בעובדה שאלגוריתם זה יכול להוות חלופה לאלגוריתמים של למידה עמוקה [14].

# עצי החלטה

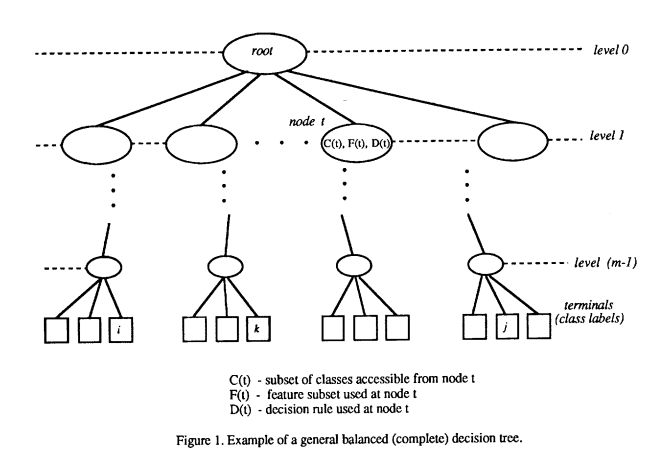
אבן הבניין הבסיסית של היער העמוק הינו מסווג הנקרא עץ החלטה. פרק זה מספק סקירה של הלוגיקה, אופן פעולתו, ומגבלותיו של עץ החלטה.

עץ החלטה הוא רצף של כללי החלטה המאורגנים בצורה היררכית, אשר ניתן לייצגם בדיאגרמת עץ [1]. הרעיון המארגן העומד בבסיסו של עץ החלטה הוא שניתן לקבל החלטה מורכבת על-ידי איחוד של מספר החלטות פשוטות [2].

מסווג זה נופל תחת קטגוריית הלמידה המפוקחת – האלגוריתם לומד לחזות משתנה כלשהו (יהא זה משתנה קטגוריאלי או רציף) על-ידי תהליך אימון, שבו ערכו של המשתנה התלוי[[2]](#footnote-2) ידוע מראש עבור כל התצפיות (תצפיות מתויגות). ערכי האמת של המשתנה התלוי משמשות כ"מורה" עבור האלגוריתם, שכן הוא יכול לתת חיזויים ולראות עד כמה הוא קרוב לערכי האמת, על-סמך פונקציית שגיאה כלשהי. תהליך האימון הוא למעשה התהליך שבו האלגוריתם עושה מינימיזציה לפונקציית השגיאה, דהיינו לומד להתקרב לערכי האמת. בתהליך הלמידה עץ ההחלטה גוזר קשרים סטטיסטיים בין המשתנים הבלתי תלויים והמשתנה התלוי. קשרים אלו מתורגמים לכללי ההחלטה של המסווג.

מסווג עץ ההחלטה הוא בעל מבנה פשוט, המורכב משורש, צמתים פנימיים ועלים. עבור תצפית חדשה שאינה מתויגת, החיזוי נקבע על-ידי המסלול שלה בעץ. בכל צומת פנימי החל מהשורש, תוצאת כלל ההחלטה (שנלמד בזמן האימון) קובעת מיהו הצומת הבא שאליו תתקדם התצפית. באופן זה התצפית ממשיכה להתקדם במורד העץ עד אשר היא מגיעה לעלה (סוף המסלול) – שם מתקבל החיזוי הסופי עבור תצפית זו.

#### איור 1: מסווג עץ החלטה [2]



לסיכום, עץ החלטה הינו דיאגרמה, בה כל צומת פנימי (כולל השורש) מציין בדיקה של כלל החלטה עבור משתנה או משתנים מסוימים, וכל ענף מייצג את תוצאת כלל ההחלטה. כאמור, כללים אלו המשמשים לחיזוי המשתנה התלוי נגזרים בשלב האימון, בו נלמד הקשר בין ערכי המשתנים הבלתי תלויים והמשתנה התלוי. לבסוף, כל עלה בעץ מייצג את החיזוי לערכו של המשתנה התלוי [1].

מצד אחד, פשטותו של מסווג עץ ההחלטה מעניקה לו יתרונות של מהירות ופרשנות. כללי ההחלטה שנלמדו יכולים לספק תובנות חשובות על היחס בין המשתנים הבלתי תלויים והמשתנה התלוי. מצד שני, פשטותו של עץ ההחלטה דווקא מגבילה אותו ולכן הוא אינו מתאים לבעיות מסוימות. למשל, עצי החלטה דורשים מבנים עמוקים כאשר מידת הקרדינאליות של המשתנה התלוי גבוהה (משתנה קטגוריאלי שיכול לקבל טווח רחב של ערכים). דרישה מינימאלית היא שמספר העלים יהיה לפחות כמספר הרמות השונות של המשתנה התלוי – שכן אחרת לא נקבל הפרדה בין רמות מסוימות. מכיוון שקיים יחס ישיר בין מספר העלים בעץ לבין עומקו, יכיל העץ הרבה מסלולים עמוקים. זהו מאפיין אשר פוגם ביכולת ההכללה (היכולת לחזות את ערך המשתנה התלוי עבור תצפיות חדשות). ככלל, ככל שהעץ עמוק יותר כך כללי ההחלטה "מותאמים" טוב יותר לתיוגן של התצפיות בשלב האימון, עד כדי כך שהכללים שנלמדו בשלב האימון מותאמים גם לרעש[[3]](#footnote-3) שקיים באצוות האימון (סט התצפיות המתויגות ששימש את המודל בשלב האימון) ולכן כללים אלו לא תקפים עבור תצפיות חדשות [6]. בעייה זו ידועה בשם overfit, או התאמת-יתר. ההבדל בין למידה והתאמת-יתר מקביל להבדל בין למידה ושינון: בעוד למידה מובילה ליכולת הכללה (בהנחה והלמידה הייתה איכותית), שינון מספק יכולת הכללה נמוכה מאוד, אם בכלל. קיימים חסרונות נוספים למסווג עץ ההחלטה כגון השפעה משמעותית של תצפיות קיצוניות (outliers), ויכולת נמוכה להתמודד עם ערכים חסרים[[4]](#footnote-4) [2].

כיום ישנן הרחבות שונות לעצי החלטה, ביניהן היער האקראי, אשר מספק מענה לבעיות שהוזכרו לעיל.

# יער אקראי

יער אקראי הוא אחד מיחידות העיבוד הבסיסיות של היער העמוק. הפרק הנוכחי נועד לספק סקירה של אלגוריתם היער העמוק – הרעיון העומד בבסיסו, מבנהו, ויתרונותיו היחסיים.

לפני שנפצח בתיאור אלגוריתם היער האקראי, נסקור שני מושגים חשובים.

## Bootstrap

המשמעות של המונח Bootstrap היא דגימה אקראית עם חזרה (כל אובייקט יכול להידגם יותר מפעם אחת). כאשר בידינו אצווה של תצפיות מתויגות, אנו יכולים לדגום חלק יחסי מן האצווה בטכניקת ה-bootstrap. ביצוע של ***k*** דגימות עם חזרה מתוך אצווה נתונה יספק לנו ***k*** תתי-קבוצות של תצפיות, כאשר קיימת שונות מסוימת בין כל תת-קבוצה מכורח אקראיות הדגימה.

ניתן להגדיל או להקטין את השונות בין על-ידי הגדרת גודל הקבוצה. למשל, ניתן להגדיר כי כל תת-קבוצה תהיה בגודל של עשרה אחוז מכמות התצפיות באצווה המקורית. כך נקבל שונות גבוהה יחסית בין תתי-הקבוצות. לעומת זאת, נוכל להגדיר שכל תת-קבוצה תהיה בגודל של תשעים אחוזים מכמות התצפיות באצווה המקורית ואז השונות בין תתי-הקבוצות תהיה קטנה יותר, והן אף יהיו דומות מאוד לאצווה המקורית.

## Bagging

המונח Bagging הוא למעשה קיצור של - Bootstrap Aggregation. ברמה הכללית ביותר, מדובר בשיטה שתכליתה לשלב גרסאות שונות של מסווג לקבלת מסווג מצרפי. היתרון של מסווג כזה הוא ניצולת של המידעים השונים (מהגרסאות השונות) על-מנת להפיק חיזוי יותר מדויק ופחות רגיש להתאמת-יתר [3].

גרסאות שונות של המסווג מתקבלות כתוצאה מהזנת תצפיות מתויגות שונות (לפחות באופן חלקי) אל אותו אלגוריתם סיווג. תצפיות אלו משמשות כ"מורה" עבור המסווג בשלב האימון, וכך השונות בין אצוות התצפיות המוזנת לכל מסווג מייצרת שונות בין כללי ההחלטה של כל מסווג (ההיגיון הפנימי של האלגוריתם על-פיו מתבצע התיוג). את השונות בין אצוות האימון המוזנות אל כל מסווג משיגים על-ידי שיטת הדגימה האקראית עם חזרה (bootstrap) שדנו בה לעיל.

המונח bootstrap, אפוא, מתייחס לאופן שבו משיגים את הגרסאות השונות של המסווג – על-ידי דגימה עם חזרה של אצוות תצפיות המשמשת את אלגוריתם הסיווג בשלב האימון. המונח aggregation מתייחס לשילובם של הגרסאות השונות של המסווג אל מסווג מצרפי. שילוב זה יכול להתבצע במספר אופנים, כאשר המקרה השכיח הוא מיצוע ערכי החיזויים כאשר מדובר בבעיית רגרסיה (חיזוי ערכו של משתנה רציף) או הצבעת רוב[[5]](#footnote-5) כאשר מדובר בבעיית סיווג (חיזוי ערכו של משתנה קטגוריאלי) [3].

## אלגוריתם היער האקראי

במאמר פורץ דרך, הציג L. Breiman את אלגוריתם היער האקראי [5]. על-אף שהיסודות למימוש יער אקראי הונחו כבר כמה שנים טרם פרסומו, השילוב בין יסודות אלו היה מהפכני. אלגוריתם היער האקראי הוא דוגמה קלאסית לכך שהשלם גדול מסכום חלקיו.

יער אקראי הוא אינו אלא אוסף של עצי החלטה המכיל אלמנט של אקראיות. פעולתו הסכמטית של היער העמוק נחלקת באופן גס לשני שלבים. ראשית, בהינתן אצווה של תצפיות מתויגות (מעתה נתייחס אל קבוצה זו בתור סט האימון השלם), מתבצעת דגימה אקראית (bootstrap) של חלק יחסי מסט האימון השלם ל ***k*** תתי-קבוצות . כל תת-קבוצה מוזנת לעץ החלטה שונה, כך שאין שני עצים שהוזנו באותה תת-קבוצה. שנית, בשלב האימון, כל עץ החלטה גוזר כללי החלטה בהתאם לתת-הקבוצה הייחודית שעליה אומן.

כאמור, מכיוון ששלב האימון עבור כל עץ החלטה ביער מתבצע על תת-קבוצה אקראית של סט האימון השלם, נוצרת שונות בין כללי ההחלטה שמפיקים העצים השונים ביער. לא רק זאת, בשביל להגדיל אף יותר את השונות בין עצי ההחלטה ביער, קיים תהליך אקראי נוסף בשלב הלמידה הקרוי random feature selection. עבור כל צומת פנימי בעץ ההחלטה, נבחרים באקראי ***n*** משתנים בלתי תלויים ומהם מנסים להפיק את כלל ההחלטה הטוב ביותר. כך למשל, נבחר המשתנה הבלתי תלוי , ומתבצעת בדיקה המכמתת את המידע שניתן להפיק מ אודות המשתנה התלוי. לאחר מכן דוגמים את המשתנים ומבצעים את אותה הבדיקה עבור כל אחד מהם. לבסוף בוחרים ב שמפיק את המידע הרב ביותר עבור המשתנה התלוי[[6]](#footnote-6). חשוב לציין ש - ***n*** הינו קטן ממש ממספר המשתנים הבלתי תלויים. אחרת, בכל צומת פנימי נבחר את כלל ההחלטה על-פי המשתנה הבלתי תלוי ה"טוב" ביותר מבין כל מרחב המשתנים הבלתי-תלויים, מה שעלול להוביל להתאמת-יתר [5].

כאשר רוצים לחזות תצפית חדשה בעזרת יער אקראי, התצפית מוזנת אל כל אחד מעצי ההחלטה ביער. תהליך החיזוי מתבצע בכל עץ בנפרד – התצפית מתחילה את דרכה בשורש העץ, ומסלולה נקבע על-פי תוצאות כללי ההחלטה בכל צומת פנימי עד אשר מגיעים לעלה, שם נקבע החיזוי. בסוף התהליך, מכיוון שמספר הערכים המנובאים עבור המשתנה התלוי הוא כמספר העצים ביער, מבוצע תהליך שקלול. שקלול סך החיזויים ביער האקראי נעשה על-פי רוב על-ידי מיצוע או על-פי הצבעת רוב (בהתאם לאופי המשתנה התלוי, רציף או קטגוריאלי). תוצאת השקלול היא זו המשמשת בתור החיזוי הסופי לתצפית זו.

לסיכום, יער אקראי הינו מצרף של עצי החלטה עם אלמנטים אקראיים. האקראיות באה לידי ביטוי בשני אופנים. ראשית, כל אחד מעצי ההחלטה ביער מאומן על קבוצה אחרת של תצפיות מתויגות. שנית, המשתנים הבלתי תלויים המעורבים בכללי ההחלטה נדגמים אקראית בכל צומת. האקראיות יוצרת שונות בין עצי ההחלטה המרכיבים את היער, ומטרתה לשפר את איכות הלמידה של המודל.

## יתרונות היער האקראי

הנחת מוצא חשובה שעומדת בבסיס היער האקראי היא ששימוש ב bagging נוטה לשפר את איכות החיזוי של המסווג. למעשה, איכות השיפור פרופורציונאלית לכמות הרעש בנתונים – ככל שכמות הרעש גדלה, כך ניכר השיפור [3].

אחת מהטענות במאמרו של Breiman היא כי שילוב של bagging יחד עם בחירה אקראית של משתנים (random feature selection), על-פיהם נלמדים כללי ההחלטה בעץ, משפרת את הביצועים של יער אקראי [5]. שיפור הביצועים קשור לכך ש bagging מקטין רגישות להתאמת-יתר, משום שכל גרסת מסווג גוזרת כללי החלטה רק על-סמך חלק (תת קבוצה ממש) מסט האימון השלם. בנוסף, random feature selection מספק נדבך נוסף של אקראיות שמספק אפילו יותר הגנה כנגד התאמת-יתר ומייצר שונות גדולה יותר בין העצים השונים ביער.

סגולתו השנייה של bagging קשורה בהערכה אד-הוק ליכולת ההכללה של המסווג. אחד מהיתרונות הבולטים של שימוש ב bagging הוא העובדה שקיימות תצפיות מתויגות שאינן נכללות בסט האימון עבור אחד או יותר מעצי ההחלטה שנמצאים ביער. כאמור, עבור כל עץ, אנו דוגמים רק חלק יחסי ממאגר התצפיות המתויגות הזמין לנו. לכן, החלק היחסי שנותר יכול לשמש לשם הערכת הטעות. ניתן לחזות את ערכי המשתנה התלוי על תצפיות אלו (אשר קרויות OOB, קיצור של Out Of Bag) ובכך לקבל הערכה של יכולת ההכללה של המודל.

קיים דמיון רב בין אלגוריתם היער האקראי וחכמת ההמונים. כל אדם, כאשר נשאל שאלה, עונה עלייה על-ידי שימוש במערכת שיקולים אינדיבידואלית שנשענת בין היתר על אמונות, ניסיון ואינטואיציות. באופן אנלוגי, האקראיות של אלגוריתם היער האקראי מייצרת "מערכת שיקולים" (כללי החלטה) אינדיבידואלית עבור כל עץ החלטה ביער. באופן כללי, חכמת ההמונים טוענת כי אם ניקח את התשובה הנפוצה ביותר (או את ה"תשובה הממוצעת") בקרב האנשים שנשאלו ונשתמש בה בתור התשובה הסופית שלנו, תהא זו תשובה טובה יותר מאשר שימוש בתשובתו של אדם בודד (כלשהו) מתוך מאגר הנשאלים. באופן דומה, יער אקראי משתמש בהצבעת רוב או מיצוע בשביל לנצל את "חכמת ההמונים" של העצים ביער.

## גרסאות שונות

הצלחתו של היער האקראי עוררה הדים והובילה לפיתוח גרסאות שונות במטרה לשפרו. אחת מן הגרסאות הללו היא יער אקראי לחלוטין ]16].

ההבדל העיקרי בין יער אקראי ויער אקראי לחלוטין נעוץ באופן קביעת כללי ההחלטה של העץ. כאמור, ביער אקראי, בכל צומת פנימי בעץ דוגמים באופן אקראי תת-קבוצה ממש של המשתנים הבלתי תלויים ומהם מנסים להפיק את כלל ההחלטה הטוב ביותר. לבסוף, המשתנה הנבחר (ובהתאם כלל ההחלטה שנקבע) הוא זה שנותן את המידע הרב ביותר על המשתנה התלוי. לשם המחשה ניתן דוגמה מפורטת כיצד זה נעשה:

נניח שעלינו לחזות האם מדובר בגבר או אישה, על סמך (1) משלח יד (2) גובה ו - (3) אורך שיער. בשלב ראשון, בצומת פנימי כלשהו בעץ נדגמים באקראי המשתנים גובה ואורך שיער. עבור כל אחד מהמשתנים שנדגמו, מתבצעת בדיקה של כללי החלטה אפשריים. למשל, עבור גובה, נבדק הכלל “***גדול מ 180 ס"מ***”, כך שכל מי שגבוה מ 180 ס"מ נשלח למסלול אחד ומי שנמוך מ 180 ס"מ נשלח אל המסלול השני. אותו תהליך בדיוק מתבצע עבור סדרה של גבהים שונים . מתבצע ניתוח של הקבוצות שמתקבלות בכל מסלול לאחר יישום הכלל – הכלל הסופי שייבחר יהא זה המייצר הפרדה מקסימאלית בין גברים ונשים. לצורך העניין, עבור הגובה 180 ס"מ, סביר להניח כי רוב האנשים שמצויים מעל גובה זה הם גברים. אי לכך, כלל החלטה זה הוא בעל ערך עבור חיזוי המשתנה התלוי. מאידך, עבור הגובה 140 ס"מ, ההבדל בין מספר הגברים ומספר הנשים מעל או מתחת לגובה זה צפוי להיות קטן. דהיינו, כלל זה אינו תורם רבות לחיזוי המשתנה התלוי. בסופו של דבר, כלל ההחלטה מייצג נקודת חלוקה של המשתנה, כזו המספקת חלוקה לקבוצות שבהן יש העדפה סטטיסטית לערך מסוים[[7]](#footnote-7) של המשתנה התלוי.

יער אקראי לחלוטין, בניגוד ליער אקראי, מייצר את כללי ההחלטה בצורה אקראית. בעוד שיער אקראי בוחן סדרת נקודות חלוקה ובוחר את נקודת החלוקה המספקת את המידע הרב ביותר עבור המשתנה התלוי, יער אקראי לחלוטין בוחר נקודת חלוקה אקראית . במבט ראשון, נדמה כי מידת האקראיות בשיטה זו רבה עד-לכדי פגימה באיכות הלמידה של המודל. אמנם, מבחנים אמפיריים מראים כי תחת תנאים מסוימים יער אקראי לחלוטין מפיק ביצועים טובים מאלו של היער האקראי [16].

# ארכיטקטורת עמוקות

פרק זה נועד לבאר את את המושג למידה עמוקה. דגש מיוחד ניתן לתיאור הארכיטקטורה של אלגוריתמי למידה עמוקה, שכן ארכיטקרטורה זו משמשת את אלגוריתם היער העמוק.

## מבוא ללמידה עמוקה

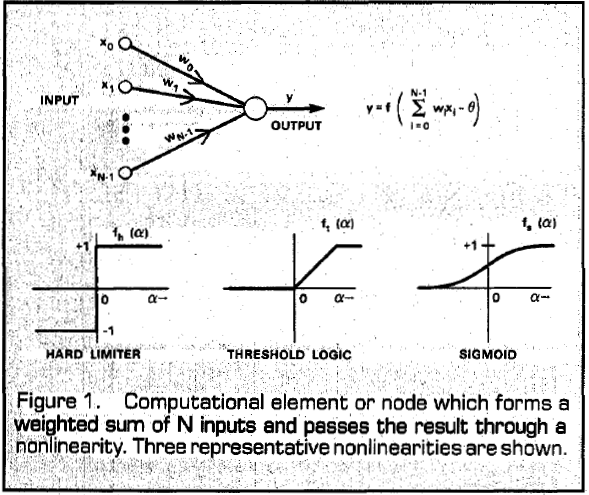
מודלים של למידה עמוקה מייצגים מערכת קשרים צפופה של יחידות עיבוד (או יחידות חישוביות). יחידות העיבוד מאורגנות יחדיו בטופולוגיה רשתית, המזכירה את הארכיטקטורות הביולוגיות של נוירונים במערכת העצבים [8]. על סמך ההקבלה הביולוגית, מודלים של למידה עמוקה מכונים רשתות נוירונים (או רשתות נוירונים מלאכותיות), כאשר נוירון הוא הכינוי של יחידת עיבוד ברשת.

טופולוגיה של רשת נוירונים ניתנת לייצוג על-ידי דיאגרמה המתארת את הקשרים השונים ברשת. באופן טיפוסי דיאגרמה זו מחולקת לשכבות, כך שהשכבה הראשונה מייצגת את הקלט הזורם אל הרשת ואילו השכבה האחרונה את הפלט היוצא ממנה. יחידות העיבוד[[8]](#footnote-8) ברשת מצויות באינטראקציה (מקיימות קשרים) על סמך הקשרים המוגדרים בארכיטקטורת הרשת. כל יחידת עיבוד היא עצמאית מחד (דבר המאפשר מקביליות חישובית), אך בעלת השפעה על הפלט הכללי באמצעות אינטראקציות עם יחידות עיבוד אחרות מאידך. הקשרים השונים בין יחידות העיבוד מתבטאים בצורה של משקולות, המעידות על טיב או חוזק הקשר. לשם המחשה, נחשוב על קבוצת אנשים שמקבלת החלטה כלשהי בארגון. כל אדם המעורב בקבלת ההחלטה הוא עצמאי, שכן את החלטתו הוא מקבל על סמך קלט מסוים ושיקולים אישיים או ארגוניים כאלה ואחרים. אמנם, מכיוון שמקבלי ההחלטות מקיימים ביניהם אינטראקציות, כל אדם הן משפיע והן מושפע מהקלט שהוא מקבל מהאנשים האחרים. בסופו של דבר, ההחלטה הכללית היא שקלול ההחלטות של כל המעורבים בדבר, כאשר כל אדם יכול להשפיע או להיות מושפע במידה כזו או אחרת מהחלטתו של אדם אחר.

רשתות נוירונים מלאכותיות ניתנות לאפיון על ידי שלושה אלמנטים: טופולוגיית הרשת, פונקציית האקטיבציה (שפעול) של יחידות העיבוד ברשת, וכללי הלמידה של הרשת [8]. נרחיב בקצרה על כל אחד מהאלמנטים. הטופולוגיה של רשת הנוירונים מתייחסת למפת הקשרים ברשת. זוהי למעשה דיאגרמה המפרטת בצורה דטרמיניסטית את המסלולים האפשריים בין כל צמתי הגרף. מדובר בגרף מכוון, כלומר, יש משמעות לכיוון הקשר. פונקציית אקטיבציה הינה פונקציה לא לינארית המיועדת לווסת את הפלט של יחידת העיבוד. פונקציות אקטיבציה שונות יוצרות יחידות עיבוד עם אופי פעולה שונה. על-ידי בחירה שונה של פונקציית אקטיבציה ניתן להשפיע על עוצמת האות, שהוא פלט יחידת העיבוד, העובר ברשת. כללי הלמידה של הרשת מתייחסים לכללים המגדירים את טיב הקשרים, כמו גם את אופי עדכון הקשרים, בין יחידות העיבוד השונות. לצורך העניין נניח שברשת נוירונים מלאכותית ישנן ***N*** יחידות קלט, כאשר כל יחידת קלט מקבלת ערך של משתנה בלתי תלוי אחר (יחידות הקלט בלתי תלויות). בנוסף, נניח כי מתוך***N*** המשתנים הזורמים כקלט לרשת, רק ***k*** משתנים אכן רלוונטיים לצורך ניבוי המשתנה התלוי ושאר המשתנים מייצגים רעש גרידא. בשלב האימון, הרשת לומדת לתת דגש רב יותר ל – ***k*** המשתנים הרלוונטיים, ולכן הקשרים בין ***k*** משתנים אלו אל יחידות העיבוד יתחזקו. חיזוק או דעיכה של הקשר מתבצעים הלכה למעשה על-ידי שינוי ערכה של המשקולת המייצגת את הקשר. ערכי המשקולות ההתחלתיים והאופן שבו יעודכנו המשקולות אלו הם למעשה כללי הלמידה של האלגוריתם.

**איור 2** מדגים את רשת הנוירונים הפשוטה ביותר, הבנויה אך ורק משכבת קלט ויחידת עיבוד יחידה המייצרת את הפלט.

#### איור 2: רשת נוירונים מלאכותית עם יחידת עיבוד אחת [8]

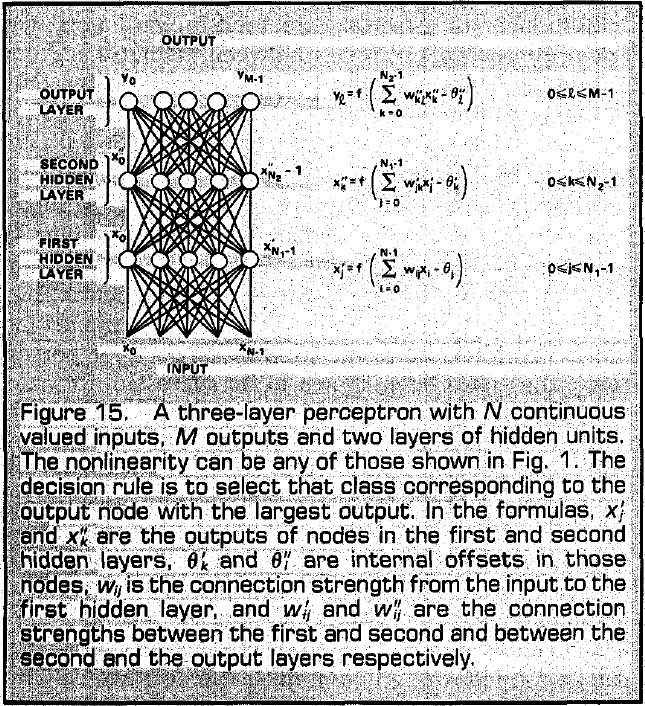


המידע ברשת זורם מיחידות הקלט ( ) אל יחידת העיבוד Y. יחידת העיבוד מבצעת שקלול של ערכי הקלט והמשקולות ( ). השקלול המודגם באיור לעיל הינו מכפלה סקלרית של המשקולות וערכי הקלט. הערך המתקבל מוזן אל פונקצית השפעול של יחידת העיבוד, מה שמייצר את הפלט סופי של הרשת.

שלושה סוגים שונים של פונקציות שפעול מתוארות באיור. נתאר בקצרה כל אחת, משמאל לימין. פונקציית השפעול הראשונה מייצרת פלט של אחד או מינוס אחד בהתאם לסף מסוים. דהיינו, עבור קלטים נמוכים דיים, הפלט הוא לעולם מינוס אחד. אולם, כאשר הקלט לפונקציית השפעול עובר סף מסוים (הסף הוא אפס במקרה הנוכחי) הפלט הוא לעולם אחד. פונקציית השפעול השנייה מתארת מצב התחלתי שבו הפלט הינו אפס עבור קלטים נמוכים דיים, אך מעל סף מסוים קיים קשר לינארי בין ערך הקלט לפלט. לבסוף, ישנו מצב של רוויה, שבו הפלט הוא לעולם אחד, על אף שערכי הקלט יכולים להיות גדולים כרצונם כל עוד הם מצויים מעל לסף הרוויה. פונקציית השפעול השלישית מתארת קשר חד-חד ערכי בין הקלט של פונקצית השפעול לפלט.

המושג למידה עמוקה קשור לכמות השכבות ברשת הנוירונים המלאכותית, שכן ככל שיש יותר שכבות כך עומקה של הרשת גדל. למשל, אלגוריתם הרגרסיה הלוגיסטית הינו בדיוק רשת נוירונים בעלת ארכיטקטורה כמתוארת ב**איור 2** עם פונקציית שפעול שונה. אולם, רגרסיה לוגיסטית אינה נופלת תחת הקטגוריה של למידה עמוקה מכיוון שזוהי רשת עם שכבה אחת ויחידה של יחידות עיבוד (רשת לא עמוקה). דיי בכך שנוסיף שכבה נוספת של יחידות עיבוד לרגרסיה לוגיסטית לקבלת רשת שמאופיינת בלמידה עמוקה.

#### איור 3: רשת נוירונים מלאכותית עם 3 שכבות [8]



לשם השוואה, הרשת המוצגת **באיור 3** אכן נופלת תחת הקטגוריה של למידה עמוקה שכן היא מכילה שלוש שכבות שונות. רשת זו מסוגלת ללמוד יחסים מורכבים מאוד בין המשתנים הבלתי תלויים. בעקבות האופי הרב-שכבתי של הרשת, נוצרים ייצוגים חדשים ומופשטים יותר של המשתנים הבלתי תלויים בכל שכבה[[9]](#footnote-9). רשת נוירונים אם כך, לא רק לומדת קשרים מורכבים בין נתוני הקלט בצורתם הגולמית, אלא גם קשרים בין ייצוגים שונים ומופשטים יותר של נתוני הקלט הגולמיים. המורכבות הגדולה שנוספת ליכולת הלמידה של הרשת המוצגת ב**איור 3** נובעת באופן ישיר מהרחבת הארכיטקטורה של הרשת החד-שכבתית שהוצגה קודם לכן. ההרחבה באה לידי ביטוי בשני דרכים: יחידות עיבוד רבות יותר בכל שכבה והגדלת מספר השכבות.

## שימושים נפוצים

רשתות נוירונים עמוקות יצרו פריצת דרך בעיבוד תמונה, וידאו, שפה, קבצי קול וטקסט [10]. למידה עמוקה הביאה עימה ביצועים יוצאי דופן במגוון תחומים ועל-כן יצרה מהפכה בעולם של למידת מכונה והתפשטה כאש בשדה קוצים.

החידוש של רשתות נוירונים מלאכותיות הוא ביכולתן האינהרנטית ללמוד ייצוגים מורכבים ומופשטים יותר של הנתונים הגולמיים [9]. מאפיין זה נקרא לימוד ייצוגיות (representation learning), והוא אינו קיים בטכניקות של למידת מכונה קונבנציונאלית ללא הנדסה ידנית מתוחכמת של המידע הגולמי. בכדי להדגים את כוחה של למידת ייצוגיות נבחן מטלת זיהוי אובייקט בתמונה, עבור רשת תלת-שכבתית:

תמונה היא לא יותר מאשר רצף של פיקסלים הניתנים לייצוג בעזרת מערך או מטריצה של ערכים מספריים. כל ערך מספרי במערך מתאר את צבעו של הפיקסל. במטלת זיהוי אובייקט, מערך הפיקסלים מוזן לרשת בצורתו הגולמית ומגיע אל שכבת העיבוד הראשונה. באופן טיפוסי, יחידות העיבוד בשכבה הראשונה לומדות נוכחות או היעדר של קונטורים (קווי מתאר). הפלט מן השכבה הראשונה נקלט על-ידי השכבה השנייה, אשר לומדת תבניות ארגון שונות של הקונטורים, קרי צורות גיאומטריות פשוטות. השכבה השלישית לומדת שילובים שונים של הצורות הגיאומטריות שנלמדו בשכבה הקודמת. המידע המעובד בשכבה השלישית מורכב דיו בשביל לייצג אובייקטים בתמונה ועל-כן עונה על דרישות המטלה.

אכן, רשתות נוירונים מלאכותיות זכו להצלחה מסחררת במטלות של עיבוד תמונה. במטלות זיהוי אובייקטים למשל, הן הביאו לשיפור הביצועים עד כדי הקטנת נפח הטעויות בחצי לעומת אלגוריתמים אחרים [12]. גם עבור עיבוד שפה וטקסט, רשתות נוירונים סיפקו ביצועים שעוד לא נראו כמותם והציבו רף חדש בעולם למידת המכונה [11]. מדובר ברשתות עם כללי למידה מתוחכמים אשר מעבדות מידע באופן סדרתי, אלמנט אחר אלמנט. החידוש ברשתות אלו הוא ביכולתן לגזור קשרים מרחביים בין האלמנטים כך שמידע על ההיסטוריה של הסדרה נשמר ברשת [10]. הדבר שימושי מאוד בעיבוד שפה שכן במטלות כמו חיזוי המילה הבאה במשפט יש משקל רב למילים שהופיעו לפנייה.

עקב הישגים יוצאי דופן, התעשייה מיהרה לאמץ את החידושים של עולם הלמידה עמוקה. כיום, כל חברות הענק כגון גוגל, באידו, ופייסבוק כולן משתמשות וממשיכות לפתח מערכות המבוססות על למידה עמוקה [12].

## מגבלות

למידה עמוקה הביאה לשינוי עמוק בעולם למידת המכונה. אולם, עדיין קיימות מגבלות המהוות מכשול מהותי במקרים רבים.

המגבלה החריפה ביותר היא כפי הנראה הצורך בנפח גדול של נתונים מתויגים [13, 14]. על-פי רוב, יישומים של למידה עמוקה מתבצעים עבור עיבוד תמונה או שפה, ולכן עלות תיוג הנתונים יקרה מאוד. למשל, עבור זיהוי אובייקט בתמונה נדרש מעבר ידני על התמונות ושיום האובייקט עבור כל תמונה ותמונה. היום, כאשר מידע נאסף באופן נמרץ ובקצב מסחרר הבעיה איננה נעוצה בנפח הנתונים הזמין (כפי שהיה בעבר) - אלא בנפח הנתונים המתויגים הזמינים, הנחוצים לצורך אימון אלגוריתמי למידה עמוקה.

המגבלה השנייה, הבולטת במיוחד לאור השוואה לאלגוריתמים אחרים, היא הכמות הגדולה של הפרמטרים האלגוריתמיים הדורשים כיוונון [14]. פרמטרים אלו כוללים את בחירת כללי הלמידה של הרשת, מספר השכבות, מספר יחידות העיבוד בכל שכבה, פונקציית השפעול של יחידות העיבוד ועוד. למידה עמוקה דורשת מיומנות רבה וניסיון בשביל לכוונן את הפרמטרים ביעילות, שכן מרחב האפשרויות הוא עצום.

# יער עמוק

פרק זה מתמקד באלגוריתם היער העמוק. אלגוריתם זה קושר יחדיו את כל האלמנטים שדנו בהם עד כה – החל מפשטותו של עץ החלטה ועד למורכבותה של רשת הנוירונים המלאכותית.

## אפיון מבנה היער העמוק

בפרק הקודם דנו בחשיבות למידת הייצוגיות בלמידה עמוקה. הטופולוגיה הרב-שכבתית של רשתות נוירונים מלאכותיות מאפשרת למידה של ייצוגים מורכבים ומופשטים יותר של הנתונים הגולמיים. למעשה, טופולוגיה זו היא המפתח לעליונות של למידה עמוקה על פני אלגוריתמים אחרים בתחומי עיבוד השפה והתמונה (אותם תחומים שדורשים ייצוגים גבוהים של הקלט). אלגוריתם היער העמוק נשען על תובנה זו - יער עמוק הינו אגד של עצי החלטה המאורגנים יחד בארכיטקטורה עמוקה.

ארכיטקטורת היער העמוק נקראת ארכיטקטורת המפולת. הרעיון הוא, בדומה ללמידה עמוקה, שכל שכבה בארכיטקטורה מעבדת את הקלט ומעבירה הלאה את הקלט המעובד. במילים אחרות, הפלט של כל שכבה משמש קלט לשכבה העוקבת. מספר השכבות בארכיטקטורת המפולת משפיע באופן ישיר על מידת המורכבות של הלמידה. ככל שמספר השכבות גדול יותר, כך ייווצרו ייצוגים מורכבים ומופשטים יותר של הקלט הגולמי.

יער עמוק, בניגוד לרשתות נוירונים מלאכותיות, מכיל שני סוגים שונים של יחידות עיבוד: יער אקראי ויער אקראי לחלוטין. השימוש ביחידות עיבוד מסוגים שונים אינו מקרי, שכן מטרתו להגדיל את שונות הלמידה בכל שכבה. הגדלת שונות הלמידה נקראת מגוון, ועיקרון המגוון הוכח כיעיל עבור מימושים של מסווג מצרפי [4]. כאשר אין מגוון כלל, אזי כל יחידות העיבוד זהות. במקרה זה, בהנחה והקלט לכל יחידות העיבוד זהה[[10]](#footnote-10), פלט שגוי ביחידת עיבוד אחת בהכרח גורר פלט שגוי של שאר יחידות העיבוד. דהיינו, השגיאה תהא קולקטיבית. לעומת זאת, כאשר קיים מגוון, אין הכרח שכל יחידות העיבוד ישגו באוניסון, וייתכן כי פלט שגוי של יחידת עיבוד מסוימת "יתוקן" על-ידי פלט נכון של שאר יחידות העיבוד בשכבה.

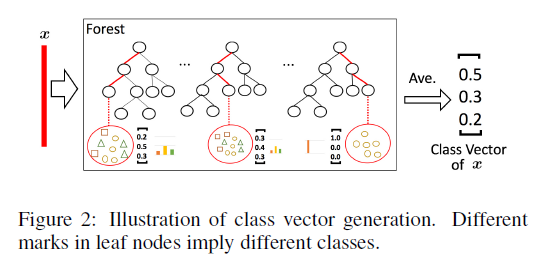
יער עמוק אם כך, מאופיין בטופולוגיה רב-שכבתית הנקראת ארכיטקטורת המפולת. מטרת ארכיטקטורה זו היא לאפשר למידה גבוהה של ייצוגיי הקלט הגולמי ובכך להפיק כללי למידה מורכבים. בכל שכבה בארכיטקטורה קיימות יחידות עיבוד משני סוגים, מה שעומד בקנה אחד עם עיקרון המגוון ונועד לשפר את ביצועי האלגוריתם.

## תיאור האלגוריתם

בסעיף זה נסקור בפרוטרוט את פעולתו של היער העמוק, החל מהקלט הגולמי ועד הפלט הסופי של המסווג. לטובת עקביות עם המקור עליו מתבסס פרק זה, נניח כי בכל שכבה של היער העמוק ישנן ארבעה יחידות עיבוד: שני יערות אקראיים ושני יערות אקראיים לחלוטין.

לשם המחשת פעולתו של האלגוריתם, נניח כי המשתנה התלוי (אותו מנסים לחזות) הוא משתנה קטגוריאלי עם שלושה ערכים אפשריים ***B***, ***A***  ו - ***C***, ונעקוב אחר המסלול שעוברת תצפית מסוימת ביער העמוק עד אשר מתקבל החיזוי הסופי לתצפית זו. ראשית, הקלט הגולמי (ערכי המשתנים הבלתי תלויים של התצפית) מועבר אל השכבה הראשונה של היער העמוק. כאמור, בשכבה זו ארבע יחידות עיבוד: שני יערות אקראיים ושני יערות אקראיים לחלוטין. הקלט הגולמי עובר דרך כל אחד מעצי ההחלטה הקיימים בכל אחד מהיערות השונים, כאשר המסלול שבו הקלט מתקדם נקבע על-ידי כללי ההחלטה האינדיבידואליים של כל עץ. לאחר שהקלט מסיים את מסלוליו (מגיע לעלה) בכל עצי ההחלטה ביער מתקבל עבורו חיזוי סופי. מכיוון שמדובר ביער (אוסף של עצים), החיזוי הסופי הינו שקלול סך החיזויים של העצים השונים ביער. תצורת החיזוי הסופי היא רשימה עם שלושה ערכים, כאשר כל ערך ברשימה מייצג את ההסתברות (מספר בטווח [0, 1]) שהתצפית שייכת לערך מסוים של המשתנה התלוי. לדוגמה, אם התקבלה הרשימה [0.2, 0.3, 0.5], כאשר מבנה הרשימה הוא [ P**(A)**, P**(B)**,P**(C)**], אזיי היער מנבא שהתצפית הנוכחית היא מסוג ***A*** בהסתברות של 0.2, סוג ***B*** בהסתברות של 0.3 וסוג ***C*** בהסתברות של 0.5. הפלט הסופי מהשכבה הראשונה ביער העמוק הוא אוסף של ארבע רשימות, כאשר כל רשימה התקבלה מיחידת עיבוד אחרת.

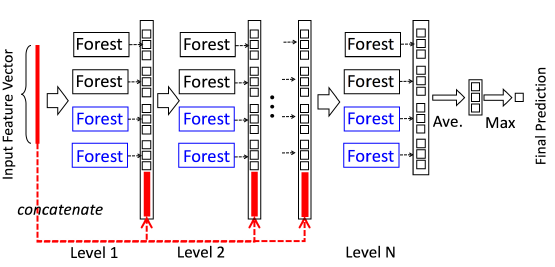
#### איור 4: יצירת רשימות החיזוי של יחידות העיבוד [14]



בשלב השני, מתבצע תהליך הנקרא stacking [7]. משמעותו הזנת פלט של אלגוריתם אחד בתור הקלט של אלגוריתם אחר[[11]](#footnote-11), רעיון הנחקר רבות בספרות. רשימות החיזוי שהתקבלו כפלט של השכבה הראשונה מוזנות בתור קלט אל השכבה השנייה של היער העמוק. בנוסף לרשימות החיזוי שהתקבלו מהשכבה הראשונה, הקלט הגולמי גם הוא מוזן אל השכבה השנייה ביער העמוק. למעשה, מתבצע שרשור של הקלט הגולמי ורשימות החיזוי מהשכבה הראשונה ליצירת קלט "מורחב" – הוא זה שנקלט על ידי יחידות העיבוד בשכבה השנייה. אם נחזור לדוגמה של קבלת החלטות בארגון ניווכח כי המצב דומה מאוד למציאות: כאשר מנהל בכיר מבצע החלטה, סביר להניח כי החלטתו מתבססת הן על הנתונים הגולמיים והן על המלצותיהם של המנהלים האחרים. המלצות המנהלים במקרה זה דומות מאוד לרשימת החיזויים שמייצרת השכבה הראשונה, שכן רשימה זו מייצגת את חוזק הנטייה ,או מידת ההמלצה, לחזות ערך מסוים של המשתנה התלוי.

הקלט המורחב שמוזן אל השכבה השנייה עובר בדיוק את אותו התהליך שהקלט הגולמי עבר בשכבה הראשונה של היער העמוק, ניתן לראות זאת ב***איור 5***. הקלט המורחב מבצע מעבר על כל עצי ההחלטה של כל יחידות העיבוד בשכבה השנייה עד אשר מתקבלות עבורו רשימות חיזוי. רשימות החיזוי הללו משמשות כפלט של השכבה השנייה והתהליך חוזר חלילה. השכבה הראשונה שונה משאר השכבות מכיוון שהקלט המוזן אל השכבה הראשונה הינו הקלט הגולמי ותו לא. לעומת זאת, בשאר השכבות מתקבל הקלט הגולמי המשורשר יחדיו עם רשימות החיזוי שהופקו על-ידי השכבה הקודמת.

#### איור 5: ארכיטקטורת המפולת [14]



בשלב האחרון, בשכבה האחרונה של היער העמוק, מתקבל החיזוי הסופי עבור התצפית. בדומה לשכבות הקודמות, גם בשכבה הסופית, כל יער מייצר רשימת ערכים המייצגת את הסתברויות החיזוי עבור כל ערך של המשתנה התלוי. על מנת להגיע לחיזוי יחיד, ראשית ממצעים את ההסתברויות מרשימות החיזויים השונים עבור כל ערך של המשתנה התלוי ולאחר מכן בוחרים את הערך בעל ההסתברות הגבוהה ביותר. לדוגמה, נניח כי רשימת החיזויים היא כדלהלן:

[0.2 , 0.2, 0.6], [0.3, 0.1, 0.6], [0.2, 0.3, 0.5] ו – [0.3, 0.2, 0.5]. על סמך רשימה זו יחושב הממוצע עבור ההסתברות שהתצפית שייכת לכל אחד מן הסוגים ***A, B*** ו – ***C****.*  ההסתברויות הממוצעות יהיו [0.25, 0.2, 0.55], מהן תיבחר ההסתברות הגבוהה ביותר, ***C*** בדוגמה הנוכחית. בהתאם לזאת, ***C*** יהיה החיזוי הסופי של היער העמוק עבור תצפית זו.

סעיף זה עסק בתיאור אופן פעולתו של היער העמוק. התיאור נעשה על-ידי התחקות אחר המסלול המלא שעוברת תצפית, החל מהשכבה הראשונה ביער העמוק ועד שמתקבל עבורה חיזוי בשכבה הסופית. כאמור, הקלט הגולמי מוזן אל השכבה הראשונה, אשר מפיקה רשימת חיזויים. רשימת החיזויים משורשרת יחד עם הקלט הגולמי ושילובם מוזן אל השכבה הבאה ביער העמוק. תהליך זה הוא רקורסיבי ונמשך עד אשר מגיעים לשכבה העליונה ביותר, בה מתקבל החיזוי הסופי. הסעיף הבא נועד להשלים את תיאור האלגוריתם על-ידי פירוט שלב האימון של היער העמוק.

## תהליך האימון

בשלב האימון, אצווה של תצפיות מתויגות מוזנת אל היער העמוק. התצפיות המתויגות משמשות ללמידת הקשרים בין המשתנה התלוי והמשתנים הבלתי תלויים, שכן ערכי האמת של המשתנה התלוי ידועים מראש. הקשרים הנלמדים מתבטאים בצורה של כללי החלטה בינאריים, כאשר מטרת תוצאת כלל ההחלטה היא לייצר את ההפרדה הטובה ביותר בין הערכים השונים של המשתנה התלוי. חשוב לציין כי הלמידה נעשית עבור כל אחת מיחידות העיבוד, יהא זה יער אקראי או יער אקראי לחלוטין, באופן עצמאי ובלתי תלוי. כלומר, ניתן להתייחס לכל יחידת עיבוד של היער העמוק כמסווג עצמאי, ואילו תפקידו של היער העמוק הינו לאגדם יחדיו בארכיטקטורה עמוקה. תהליך הלמידה של היער העמוק כולו מתבצע אם כך ברמת יחידת העיבוד – מדובר בדיוק באותו תהליך למידה אותו סקרנו עבור יער אקראי ויער אקראי לחלוטין בפרקים הקודמים.

רשתות נוירונים מלאכותיות סטנדרטיות דורשות כי ארכיטקטורת הרשת תהיה מוגדרת מראש: מספר השכבות ברשת, מספר יחידות העיבוד בכל שכבה, פונקציית השפעול ופרמטרים נוספים. הגדרת ארכיטקטורת הרשת מראש אינה מטלה טריוויאלית, שכן מספר הפרמטרים הניתנים לכיוונון הוא עצום ושוני בין ארכיטקטורות שונות יכול להוביל להבדלים משמעותיים בביצועי הרשת [14]. יער עמוק, בניגוד לרשתות נוירונים, מווסת את הארכיטקטורה שלו בהתאם למורכבות הבעיה. רוצה לומר, מורכבות הבעיה מותאמת אינהרנטית לעומק הארכיטקטורה כך שככל שמורכבות הבעיה גדלה הארכיטקטורה תהיה עמוקה יותר.

וויסות הארכיטקטורה נעשה כחלק מתהליך הלמידה של היער העמוק, בעזרת אומדן ביצועיו של המודל. אומדן הביצועים נעשה על-ידי חיזוי של המודל עבור תצפיות מתויגות שאינן מהוות חלק מאצוות הלמידה, כך שניתן יהיה להשוות את החיזויים שהתקבלו עם ערכי האמת של המשתנה התלוי. חשוב לציין כי תצפיות אלו אינן נחשפות למודל בשלב האימון, הן משמשות אך-ורק להערכת הביצועי המודל. אצווה של תצפיות כאלו קרויה אצוות אימות (בניגוד לאצוות אימון שעליהן המודל מתאמן) ומטרתן לדמות את המצב האמיתי שבו המודל מנבא ערכים עבור נתונים חדשים ולא מתויגים. לאחר שנוצרה שכבה חדשה ביער העמוק, מתבצע חיזוי על אצוות האימות על-מנת להעריך את ביצועי המודל. במידה וביצועי המודל בשכבה שנוספה לא השתפרו משמעותית בהשוואה לביצועי המודל בשכבה הקודמת לה אזיי תהליך הלמידה נפסק. אין צורך ביצירת שכבות נוספות, והשכבה האחרונה שנוספה נקבעת להיות השכבה הסופית של היער העמוק. במילים אחרות, מורכבות (עומק הארכיטקטורה, מספר השכבות ביער העמוק) היער העמוק גדלה רק כאשר יש בכך צורך והדבר תורם לביצועי המודל. בעקבות מאפיין זה היער העמוק ניתן ליישום בספקטרום רחב של בעיות, החל ממטלת חיזוי פשוטות (שאינן דורשות ייצוגים מופשטים של הקלט הגולמי) ועד לעיבוד תמונה ושפה.

לסיכום, תהליך הלמידה של היער העמוק מתבצע ברמה הלוקאלית בכל יחידת עיבוד בנפרד. תהליכי הלימוד (שלבי האימון של המודל) הינם אותם תהליכי לימוד סטנדרטיים של יער אקראי ויער אקראי לחלוטין שנסקרו בפרקים הקודמים. בנוסף, קיימת רגולציה עצמית של מורכבות היער העמוק על-ידי הערכת ביצועיו בשלב הלמידה, עבור כל שכבה שנוצרת. באופן זה, מורכבות מתווספת למודל רק במידה ויש בכך רווח משמעותי המשתקף בביצועי המודל.

## Multi-Grained Scanning

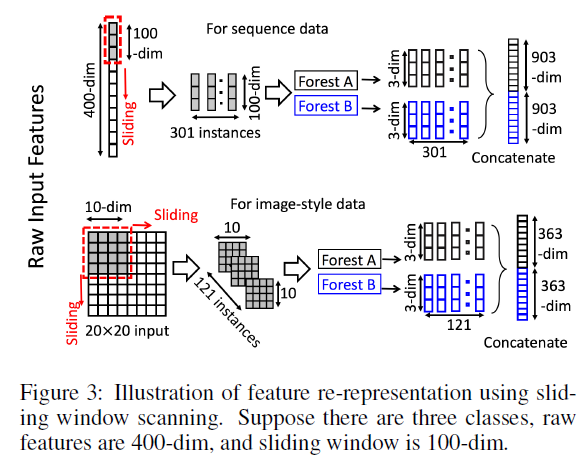
Multi-Grained Scanning היא שיטה שנועדה לסנתז משתנים מורכבים ומופשטים על-ידי סריקה שיטתית של הקלט הגולמי. בשביל להבין את הרציונאל מאחורי שיטת ה Multi-Grained Scanning הממומשת ביער העמוק, נחזור לרגע לעולם רשתות הנוירונים המלאכותיות. כזכור, המבנה הרב-שכבתי של רשתות נוירונים מלאכותיות מאפשר לרשת ללמוד ייצוגים גבוהים ומופשטים יותר של הקלט הגולמי. אף יותר מכך, הרשת מסוגלת ללמוד על קשרים ושילובים שונים של ייצוגים אלו. מאפיין זה משחק תפקיד מרכזי במטלות מורכבות כגון עיבוד תמונה ושפה. ניקח בתור דוגמה את תחום עיבוד התמונה, ונתייחס אל קו בתור ייצוג גבוה של רצף פיקסלים קהים סמוכים בתמונה. נניח, לצורך העניין, כי הפלט המיוצר על-ידי יחידות העיבוד של השכבה השנייה ברשת מקביל לקווים שונים שנמצאים בתמונה. אזיי, בשכבה השלישית של הרשת, יחידות העיבוד ילמדו שילובים שונים של קווים אלו. שילובים אלו יכולים להיות צורות גיאומטריות כגון אותיות בשפה, ספרות, חפצים פשוטים וכדומה. שכבה נוספת תגדיל עוד יותר את המורכבות ותאפשר למידת שילובים של צורות גיאומטריות שונות לכדי אובייקטים: זיהוי פנים, מספר לוחית רישוי, מילים בשלטי חוצות וכדומה. Multi-Grained Scanning זוהי אופטימיזציה שנועדה לייצר יכולת למידה דומה עבור יער עמוק, על-ידי סריקה מתודולוגית של הקלט הגולמי.

שיטת ה Multi-Grained Scanning ממומשת בשתי דרכים שונות בהתאם לאופי הבעיה. קיים מימוש עבור המקרה של מידע סדרתי (כמו עיבוד שפה), ומימוש נוסף עבור המקרה של מידע ויזואלי (עיבוד תמונה) – נתאר את שניהם. עבור מידע סדרתי, ישנה משמעות ניכרת לאלמנטים סמוכים בסדרה (למשל מילים סמוכות במשפט). לצורך הדגמת השיטה, נניח שהקלט הגולמי הינו סדרה של 400 ערכים. מוגדר חלון הזזה (sliding window) בגודל של 100 אלמנטים, שמטרתו לחלק את הקלט הגולמי לתתי-סדרות רציפות. מאה הערכים הראשונים בסדרת הקלט הגולמי יוצרים תת-סדרה ראשונה, תת-הסדרה השנייה מורכבת מהערך השני ועד הערך שמיקומו 101 בסדרה, תת-הסדרה השלישית מהערך במיקום 3 ועד הערך במיקום 102 וכן הלאה. באופן זה מתקבלות 301 תתי-סדרות, שמטרתן להבליט השפעה של אלמנטים סמוכים בסדרה. עבור המקרה של מידע ויזואלי, התהליך הכללי זהה, בעוד שאופן מימוש החלון ותזוזתו שונים. קלט של מידע ויזואלי מאורגן בצורה של מטריצת ערכים (בניגוד לסדרה רציפה של ערכים במקרה של מידע סדרתי). בהתאם לזאת, חלון ההזזה מוגדר אף הוא להיות מטריצה, בעל מימדים קטנים יותר מאלו של מטריצת הקלט. שיטת ה- Multi-Grained Scanning עבור מידע ויזואלי מחלקת את מטריצת הקלט לתתי-מטריצות, בכדי להבליט קשרים מרחביים בין האלמנטים השונים בקלט. חלון ההזזה נע בצורה אופקית, החל מהחלק השמאלי העליון של המטריצה ועד החלק הימני העליון. לאחר מכן, מתבצעת תזוזה אנכית של החלון כלפי מטה ("ירידת שורה"), ותהליך הסריקה האופקית חוזר חלילה. התהליך נמשך עד אשר מגיעים לחלק הימני התחתון של המטריצה, מה שמבטיח כי כל האזורים השונים בתמונה נסרקו באמצעות החלון. נוסיף כי תהליך חלוקת הקלט על-ידי Multi-Grained Scanning יכול להתבצע עבור חלונות בגדלים שונים. גודל החלון שולט על גרעיניות החלוקה של הקלט, ויכול להוביל לשיפור משמעותי של הביצועים, שכן בעיות מסוימות יכולות להפיק תועלת מחלוקה גסה של הקלט בעוד שבעיות אחרות דורשות חלוקה עדינה יותר.

חלוקת הקלט הגולמי לתתי-קלטים נועדה לשפר את ביצועי המודל על-ידי סריקה דקדקנית יותר של האזורים השונים במרחב הקלט בהתאם לאופי הנתונים[[12]](#footnote-12). כל אחד מתתי-הקלטים שנוצרו מוזן ליער אקראי ויער אקראי לחלוטין לקבלת רשימת חיזויים, בתבנית המוכרת של [ P**(A)**, P**(B)**,P**(C)**] (תחת ההנחה ש ***A, B, C*** הם ערכיו האפשריים של המשתנה התלוי). לדוגמה, אם הופקו בשלב ה Multi-Grained Scanning 301 תתי-סדרות של הקלט הגולמי, אזיי יהיו 301 רשימות חיזויים מהיער האקראי ו – 301 רשימות חיזויים נוספות מהיער האקראי לחלוטין. רשימות החיזויים בעקבות Multi-Grained Scanning מתייחסות, כל אחת בהתאמה, לייצוג גבוה של אזור מסוים במרחב הקלט. רשימות אלו הן למעשה המקבילות של הייצוגים המופשטים הנוצרים בשכבות עיבוד גבוהות של רשתות נוירונים מלאכותיות – הן מייצגות את הקלט הגולמי המעובד לאחר שעבר תהליך למידה כלשהו. לבסוף, רשימות החיזויים הללו משמשות כמרחב קלט גולמי חדש המשמש ללמידה של היער העמוק.

#### איור 6: Multi-Grained Scanning מייצר מרחב-קלט חדש מהקלט הגולמי

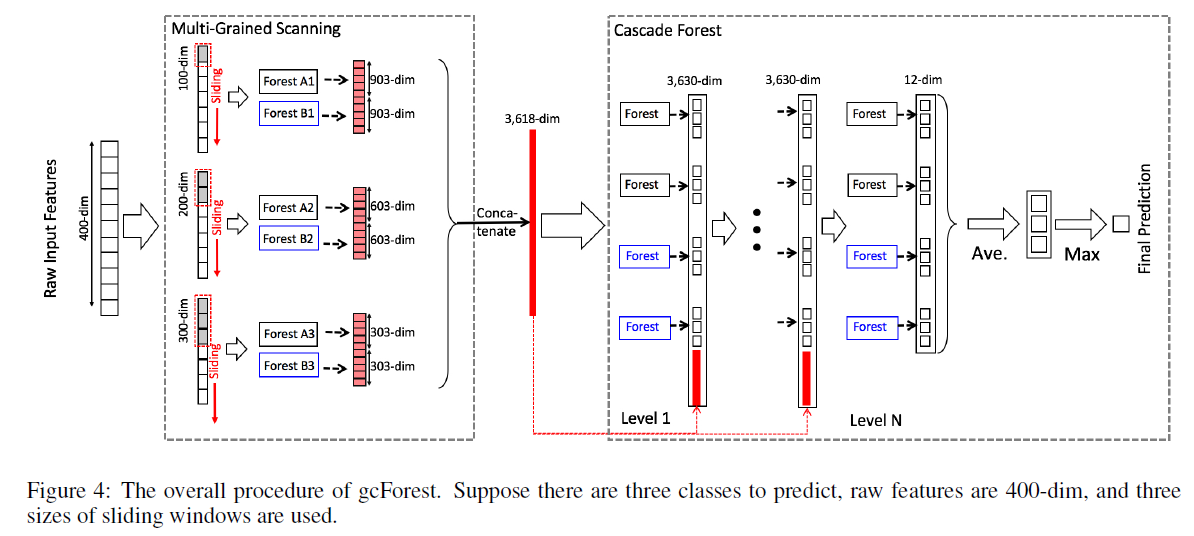
#### עבור מידע סדרתי (למעלה) ומידע ויזואלי (למטה) [14]



נחזור לדוגמה אנלוגית של קבלת החלטות בארגון, יהא זה משרד פרסום, בכדי להבין טוב יותר את הרציונאל מאחורי חלוקת הקלט לתתי-סדרות והתועלת שניתן להפיק מייצוגים מופשטים של תתי-סדרות אלו. המטרה של משרד הפרסום היא גריפת רווחים מקסימאליים מהמודעות אותן הם מפרסמים. נניח כי הדיון הוא האם כדאי לסגור מודעה כלשהי או להמשיך לפרסם אותה (מטלת חיזוי של משתנה קטגוריאלי). הקלט הגולמי – נתוני המודעה – מכיל מידע על עלות המודעה, קהל היעד שאליו מפרסמים (אזור גיאוגרפי וטווח גילאים), גודל המודעה, והרכיבים הויזואליים של המודעה (תמונה וטקסט). בדיון, משתתפות התקציבאית, הגרפיקאית, מנהלת השיווק והמנהלת הכללית. התקציבאית בוחנת בעיקר שיקולי עלות, הגרפיקאית מתמקדת במאפיינים הויזואליים של המודעה ואילו מנהלת השיווק חושבת על קהל היעד. כלומר, כל אחד מהמשתתפים בדיון ממוקד באזור מסוים אחר של מרחב הקלט. כאשר מתנהל הדיון, כל אחד מאנשי המקצוע מביע את דעתו בפני המנהלת הכללית. דעתה של כל אחת מנשות המקצוע אמנם מבוססת על הקלט הגולמי שבה בחרה להתמקד, אך מדובר בייצוג מופשט שלו, מכיוון שהקלט עבר תהליך עיבוד שבעקבותיו נחרצה דעתה של אשת המקצוע. לבסוף, דעותיהן השונות של נשות המקצוע נשמעות ונשקלות כאשר המנהלת הכללית מקבלת את ההחלטה הסופית.

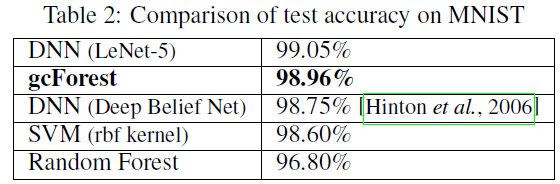
נסכם - שיטת ה Multi-Grained Scanning היא שיטת אופטימיזציה שנועדה לאפשר ליער עמוק ללמוד שילובים של ייצוגים גבוהים של הקלט הגולמי (כמו ברשתות נוירונים מלאכותיות). בשיטה זו, ראשית מחלקים את הקלט הגולמי לתתי-קלטים, בהתאם לאופי הנתונים (מידע סדרתי או מידע ויזואלי). שנית, יער אקראי ויער אקראי לחלוטין מאומנים על כל אחד מתתי-הקלטים הללו, בכדי לייצר רשימת חיזויים – ייצוג גבוה של תתי הקלטים. לבסוף, רשימות החיזויים שהופקו משורשרות יחדיו ליצירת מרחב קלט מורחב, כאשר מרחב קלט זה משמש בתור הקלט הסופי שמוזן אל היער העמוק.

#### איור 7: תהליך הלמידה המלא של יער עמוק [14]

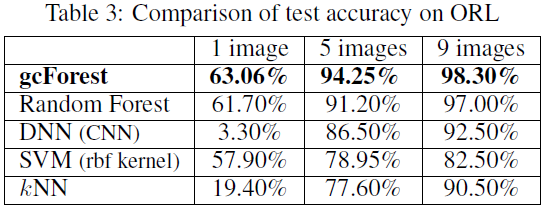


## מבחני ביצוע

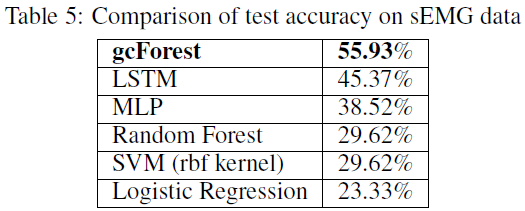
בחינה אמפירית של ביצועי היער העמוק נעשו על מקבצי נתונים מוכרים, בפרק זה נסקור כמה מהתוצאות המוצגות ב [14]. ראשית, להלן התוצאות עבור נתוני ***MNIST***[17], בהם צריך לזהות מהי הספרה הרשומה בתמונה:



שנית, עבור נתוני ***ORL*** [18] – מטלת זיהוי פרצופים. שלושה גרסאות שונות נבחנו במטלה זו, אשר נבדלות בשלב האימון. בגרסה הראשונה, תשע תמונות שונות של כל אדם נלמדו בשלב האימון. בגרסה השנייה, חמש תמונות שונות של אותו אדם, ובגרסה השלישית תמונה אחת בלבד. המטרה במטלה זו היא שיוך הפרצוף אל האדם המתאים. ההבדלים בין התמונות השונות של אותו אדם כוללים זמן צילום שונה, תאורה שונה, הבעות פנים ופרטי פנים (עם/בלי משקפיים).



לבסוף, להלן השוואה עבור מקבץ נתוני ה ***sEMG*** [19]. המטרה בנתונים אלו היא לזהות את סוג תנועת היד (שישה תנועות אפשריות סך הכול) לפי נתוני הקלטה של חיישן ה EMG. אופי הנתונים הוא סדרתי - חיישן ה EMG מקליט 500 ערכים בשנייה. להלן התוצאות:



התוצאות של מבחני הביצוע (benchmarks) השונים, מראות באופן חד-משמעי כי היער העמוק מפיק ביצועים ברי-השוואה עם אלגוריתמים מעולם הלמידה העמוקה או אלגוריתמים מובילים אחרים. במקרה הראשון של ***MNIST***, יער עמוק מספק ביצועים דומים מאוד, אם כי פחותים במעט, לאלגוריתמים שנחשבים חוד-החנית בבעיות מסוג זה. במקרה השני, עבור זיהוי פרצופים, יער עמוק מגיע לביצועים גבוהים מאוד ואף עולה על ביצועי רשת נוירונים מלאכותית (DNN בטבלה). במקרה השלישי של ***sEMG***, יער עמוק מראה עליונות ברורה ומספק ביצועים טובים בהרבה אל מול המתחרים.

## הרחבות אפשריות

ברמה הכללית, יער עמוק הינו אוסף של יחידות עיבוד המאורגנות בארכיטקטורה עמוקה. מכיוון שכל אחת מיחידות העיבוד לומדת באופן עצמאי (הפקת כללי ההחלטה מתבצעת עבור כל יחידת עיבוד בנפרד), ניתן בנקל להרחיב את היער העמוק כך שיכלול אלגוריתמים מגוונים נוספים.

מימוש היער העמוק המתואר ב [14], אותו סקרנו עד כה, מכיל שני סוגי אלגוריתמים בלבד בתור יחידות עיבוד: יער אקראי ויער אקראי לחלוטין. אולם, לא קיימת אף מגבלה המחייבת שימוש בלבדי בשני אלגוריתמים אלו. הדרישה היחידה היא, שבהינתן אצוות אימון (קבוצה של תצפיות מתויגות) אלגוריתם המשמש יחידת עיבוד ביער העמוק ילמד להפיק כללי החלטה. אי לכך, כל אלגוריתם המבוסס על כללי החלטה יכול לשמש כיחידת עיבוד ביער העמוק. לדוגמה, ניתן באופן מיידי לשלב ביער עמוק גם גרסאות אחרות של יער אקראי בתור יחידות עיבוד, כגון יער הרוטציה [15].

מבנהו הדינאמי של היער העמוק, אפוא, פותח צוהר להרחבות רבות, המתבטאות במגוון האלגוריתמים המשמשים כיחידות עיבוד. הרחבות כאלו עומדות בקנה אחד עם עיקרון המגוון [4] ויכולות להוביל לשיפור נוסף בביצועיו של היער העמוק.

# 

# סיכום

הסמינר הנוכחי נועד להציג אלגוריתם חדשני בשם יער עמוק. בעזרת תובנות מעולם הלמידה העמוקה ואלגוריתמים מתקדמים של למידה מפוקחת, יער עמוק מעיד על ביצועים מעולים במגוון רחב של בעיות, בעלות מידת-מורכבות שונה, ומעולמות תוכן שונים.

היחידה החישובית הבסיסית של היער העמוק מתבססת על עץ ההחלטה - אלגוריתם פשטני הגוזר כללי החלטה סטטיסטיים בין המשתנים הבלתי תלויים והמשתנה התלוי. במרוצת השנים, עקב מגבלותיו של עץ ההחלטה, פותחו אלגוריתמים חדשים המושתתים על שילוב של מספר עצי החלטה שונים בכדי לייצר חיזוי עבור המשתנה התלוי. אולי המפורסם ביניהם הוא יער אקראי, אשר מטרתו לייצר שונות בין עצי ההחלטה במטרה ללמוד אספקטים שונים של מטלת החיזוי ולמנוע מצב של התאמת-יתר (overfit). אלגוריתם נוסף הינו היער האקראי לחלוטין, המייצר אלמנט חזק יותר של אקראיות מאשר זה הממומש ביער אקראי. שני האלגוריתמים הללו - יער אקראי ויער אקראי לחלוטין - נבחרו שניהם (בהתאם לעיקרון המגוון) לשמש בתור היחידות החישוביות של היער העמוק.

היחידות החישוביות של היער העמוק מאורגנות בטופולוגיה רב-שכבתית. טופולוגיה זו היא הכלי שמאפשר למידת ייצוגים גבוהים ומופשטים (למידת ייצוגיות) של הקלט הגולמי, דבר בעל תועלת רבה כאשר מורכבות הקלט גבוהה (דוגמת עיבוד תמונה ועיבוד שפה). על-ידי עיבוד חוזר ונשנה של הקלט בשכבות השונות, יער עמוק מסוגל להפיק כללי למידה מורכבים מאוד, ברי-השוואה ליכולות הלמידה של רשתות נוירונים מלאכותיות. יחד עם זאת, בניגוד לרשתות נוירונים מלאכותיות, הארכיטקטורה של יער עמוק מווסתת על-ידי הערכת ביצועים תמידית בעת הלמידה. רוצה לומר, יער עמוק מתאים את הארכיטקטורה שלו למורכבות הנתונים באופן עצמאי. בנוסף, שיטת אופטיציזציה בשם Multi-Grained Scanning משפרת את רגישות היער העמוק לאזורים מסוימים, כמו גם תכונות מסוימות, במרחב הקלט. בשילוב עם למידת ייצוגיות, שיטה זו מייצרת סריקה דקדקנית של מרחב הקלט ברמות עיבוד שונות – דבר שמגדיל עוד יותר את יכולות הלמידה של יער עמוק.

יער עמוק הוא שילוב של ארכיטקטורה יעילה עם יחידות חישוב חכמות, מבנה מתוכנן-היטב עם פונקציונאליות גבוהה. המבנה והפונקציונאליות של היער העמוק פועלים בהרמוניה בשביל לייצר כלי חיזוי גמיש, המותאם למגוון רחב של בעיות. בעוד רשתות נוירונים דורשות מאגר למידה רחב (נפח גדול של נתונים מתויגים) וכוונון מדוקדק של פרמטרים, מאפייני היער העמוק מייתרים זאת כמעט לחלוטין. לא בכדי מחברי האלגוריתם מכנים אותו "חלופה לרשתות נוירונים מלאכותיות".

# מקורות

[1] Swain, P.H., and Hauska, H. 1977. **The decision tree classifier: design and potential**. IEEE *Transactions on Geoscience and Electronics, 15, 142-147.*

[2] Safavian, S. R., and Landgrebe, D. 1991. **A survey of decision tree classifier methodology**. *IEEE Transactions on Systems, Man, and Cybernetics, 21, 660–674*.

[3] Breiman, L. 1996. **Bagging Predictors**, *Machine Learning, 26, No. 2, 123-140.*

[4] Dietterich, T. G. 2001. **Ensemble methods in machine learning**. *In J. Kittler and F. Roli, editors, Multiple Classifier Systems, pages 1–15. LNCS Vol. 1857, Springer*.

[5] Breiman, L. 2001. **Random forests**. *Machine learning, 45(1):5–32*.

[6] T. Dietrich. 1995. **Overfitting and undercomputing in machine learning.** *ACM Comput. Surv.27:326–327.*

[7] Džeroski S, Ženko B. 2004. **Is combining classifiers with stacking better than selecting the best one?** *Mach Learn 54(3):255–273.*

[8] Lippmann R. P. 1987. **An introduction to computing with neural nets**, IEEE ASSP Mag., 4, 4-22.

[9] Bengio, Y., Courville, A. & Vincent, P. 2013. **Representation learning: a review and new perspectives**. *IEEE Trans. Pattern Anal. Machine Intell. 35, 1798–1828.*

[10] LeCun, Y., Bengio, Y. & Hinton, G. 2015. **Deep learning**. *Nature 521, 436–444*.

[11] Graves, A., Mohamed, A. R., & Hinton, G. 2013. **Speech recognition with deep recurrent neural**

**networks**. In 38th IEEE International Conference on Acoustics, *Speech and Signal Processing*, pages 6645–

6649.

[12] Krizhenvsky, A., Sutskever, I. & Hinton, G. 2012. **ImageNet classification with deep convolutional neural networks**. *In Advances in Neural Information*, Processing Systems 25, pages 1097–1105.

[13] Chen, X. W., Lin, X. 2014. **Big data deep learning: Challenges and perspectives**. *Access, IEEE, vol. 2, pp. 514–525*.

[14] Zhou Z. H. and Feng. J. 2017. **Deep forest: Towards an alternative to deep neural networks**. *In: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-2017)*, Melbourne, Australia.

[15] Rodriguez , J. J., Kuncheva, L. I. and Alonso C. J. 2006. **Rotation forest: A new classifier ensemble method**. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10):1619–1630*.

[16] Liu, F. T., Ting, K. M., and Zhou, Z. H. 2008. **Spectrum of variable-random trees**. *Journal of Artificial Intelligence Research 32, 355–384*.

[17] Image collection of handwritten digits: <http://yann.lecun.com/exdb/mnist/>

[18] Face images taken between April 1992 and April 1994: <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>

[19] Databases of surface electromyographic signals of 6 hand movements using Delsys' EMG System: <https://archive.ics.uci.edu/ml/datasets/sEMG+for+Basic+Hand+movements>

1. בנתונים מתויגים ידועים ערכי האמת של הערך אותו האלגוריתם מנסה לחזות. למשל עבור מטלת זיהוי ספרות מתוך תמונה נתונים מתויגים משמע שעבור כל תמונה ידוע בדיוק איזו ספרה מופיעה בתמונה. [↑](#footnote-ref-1)
2. הבחנה סינתטית בין המשתנה אותו המסווג רוצה ללמוד לחזות ושאר המשתנים המשמשים מנבאים עבורו. המשתנה המנובא "תלוי" בערכם של המשתנים האחרים ומכאן שמו. [↑](#footnote-ref-2)
3. ברוב המקרים קיימות טעויות מדידה, טעויות שנובעות מגורם אנושי, משתנים שאינם רלוונטיים למשתנה המטרה וכדומה. [↑](#footnote-ref-3)
4. סקירה נרחבת של המגבלות ניתן לראות ב – [2], יחד עם הצעות לפתרון עבור חלקן. [↑](#footnote-ref-4)
5. לצורך העניין נדמיין שישנן חמש גרסאות שונות של המסווג, שמטרתו לנבא האם ירד או לא ירד גשם ביום מסוים. תרחיש אפשרי הוא ששניים יחזו שאכן ירד גשם היום בעוד ששלושה יחזו שלא יירד גשם היום. על-פי עקרון הצבעת רוב החיזוי הסופי יהיה שלא ירד גשם. יש לשים לב שעבור מטרות סיווג כדאי להגדיר מספר אי-זוגי של גרסאות בכדי שתמיד תהיה הכרעה (יהיה רוב) עבור החיזוי הסופי. [↑](#footnote-ref-5)
6. המטריקה המדויקת לכימות המידע שמספק כל משתנה ביחס למשתנה התלוי איננה הכרחית לצורך הצגת האלגוריתם. נהוג להשתמש במדד ג'יני, אך הדיון בו חורג מהתחום של הסמינר הנוכחי. עם זאת, בסעיף ***גרסאות שונות*** (סוף הפרק הנוכחי) קיים הסבר מעמיק יותר על תהליך כימות המידע בעצי החלטה. [↑](#footnote-ref-6)
7. בדוגמה הנ"ל מדובר בחיזוי של משתנה בינארי (גבר או אישה) ומכאן הניסוח ערך מסוים. לעומת זאת, ייתכנו בעיות שבהן בעיית החיזוי היא של יותר משני ערכים, למשל חיזוי מצב משפחתי (רווק, נשוי, גרוש, אלמן). הניסוח המדויק יותר עבור המקרה הכללי הוא ערכים מסוימים, ולאו דווקא ערך מסוים (יחיד), שמתאים למקרה הבינארי בלבד. [↑](#footnote-ref-7)
8. הבחירה הסמנטית להשתמש במילה יחידת עיבוד ולא במושג נוירון נעשתה במכוון. נוירון הינו יחידת עיבוד פשוטה, והיכולות החישוביות שמתקבלות מרשת נוירונים הינן תוצר של מורכבות וחוזק הקשרים בין הנוירונים. לעומת זאת, המושג יחידת עיבוד אינו מעיד באופן לטנטי על מידת המורכבות של יחידת העיבוד - זוהי נקודת מפתח לטובת פיתוחו של אלגוריתם היער העמוק המוצג בהמשך. [↑](#footnote-ref-8)
9. הרחבה על ייצוגים מופשטים של הקלט הגולמי והתועלת בהם ניתנת בסעיף הבא. [↑](#footnote-ref-9)
10. יצירת שונות בלמידה עבור רשתות נוירונים מלאכותיות מתבססת על הפרה של הנחה זו. אמנם, כל יחידות העיבוד ברשתות נוירונים זהות, אך הקלט המוזרם אליהן אינו – קיימת אקראיות בתהליך מתן המשקולות (הגדרת חוזק הקשרים) בין יחידות הקלט ליחידות העיבוד, כך שכל יחידת עיבוד מקבלת קלט מעט שונה. [↑](#footnote-ref-10)
11. stacking מתרחש באופן אינהרנטי בארכיטקטורות עמוקות בשל המבנה הרב-שכבתי שלהן. אולם, השימוש ב - stacking אינו מוגבל רק לארכיטקטורות עמוקות, ניתן לסדר כל קבוצת אלגוריתמים כלשהי בצורה טורית כך שהפלט של האלגוריתם הראשון מתקבל כקלט לאגוריתם השני וכן הלאה. [↑](#footnote-ref-11)
12. לא קשה לדמיין מצב שבו סריקה של אזור מסוים בלבד בקלט נחוצה לפתרון הבעיה. נחשוב למשל על מודל שמטרתו לסרוק את סכום השיק לאחר שהוא מופקד בבנק דרך הכספומט. השייק אמנם נסרק בשלמותו, אך סכומו נכתב תמיד רק באזור המיועד לכך – חלק מאוד מסוים בתמונה. אם אלגוריתם החיזוי ינסה למצוא את כל רצפי המספרים בתמונה ורק לאחר מכן להחליט מהו סכום השיק, הרי שיש יותר סיכוי לטעות (בלבול בין מספר תעודת זהות לסכום השיק לדוגמה) ומתבצעת עבודה חישובית שאינה נחוצה. [↑](#footnote-ref-12)