Handwriting-Based Gender Classification

מגישים : אורי טביבי ושון אסולין

מנחות אקדמיות : ד"ר אירינה רבייב, ד"ר מרינה ליטבק

המכללה האקדמית להנדסה ע"ש סמי שמעון, באר שבע

יוני 2021

המכללה האקדמית להנדסה ע"ש סמי שמעון

מהנדסים לעולם טוב יותר!

בסביבת PROJECT ORIENTED

תוכן העניינים

זיהוי מגדר על בסיס תמונות כתבי יד

BS_SE-21-104

מנחות: ד"ר אירינה רבייב
ד"ר מרינה ליטבק

שון אסולין
Shonas@ac.sce.ac.il

אורי טביבי
Orita4@ac.sce.ac.il

תקציר

זיהוי מגדר לפי כתב יד נחשבת לאחת הבעיות הנחקרות והמעניינות המעוררת עניין רב בשל מגוון התחומים הרחב שנעשה בהם שימוש, החל מרפואה לצרכי מחקר על התנהגות ובריאות האדם ועד לשימושים נרחבים כמו אימות נתונים, ניתוח מסמכים היסטוריים וחקירות קרימינולוגיות.

פרויקט זה עוסק במחקר ופיתוח כלי אוטומטי המבוסס על רשתות נוירונים לזיהוי מגדר על בסיס תמונות כתבי יד, חוקרים רבים מראים שבעיה זו לא פשוטה לפתרון על ידי בני-אדם שלא מומחים בנושא זה.

במסגרת המחקר בחנו שימוש במספר מודלים של רשתות נוירונים עמוקות לזיהוי תמונה, המבצעים חילוץ מאפיינים מסריקות כתבי היד ולאחר מכן מזהות את מין הכותב. השימוש במודלים אלו מראה יכולות טובות לזיהוי המאפיינים המבדילים בין השוני של כתבי היד לפי מגדר.

הניסויים התבססו על מאגר HHD_gender המכיל תמונות כתבי יד בשפה העברית, כמו כן, על מנת להרחיב את המחקר ולהעריך את טיב המודלים, ביצענו ניסויים נוספים על מאגר QUWI המכיל תמונות כתבי יד בשפות האנגלית והערבית.

בנוסף, במהלך המחקר ערכנו סקר שהופץ לקהל הרחב כדי להשוות ביצועים בין זיהוי המתבצע על ידי בני אדם שלא מומחים בניתוח כתבי יד לבין שימוש ברשתות נוירונים.

בשלב השני לפרויקט, בנינו מערכת (אתר) המאפשרת למשתמש להעלות סריקה של כתב יד ולקבל ניתוח מפורט לגבי המידע שהתקבל בתהליך זיהוי המין של הכותב.

במסגרת המחקר ראינו כי קיים שוני בין כתבי היד של שני המינים וכי שימוש בכלים אוטומטיים מביא לתוצאות טובות יותר בזיהוי לעומת זיהוי הנעשה על ידי בני אדם שלא בקיאים בנושא.

במהלך המחקר נכתב מאמר אקדמי על ידי צוות הפרויקט הכולל את ד"ר מרינה ליטבק וד"ר אירינה רבייב, המתאר את המחקר והעבודה שבוצעה. מאמר זה נמצא בשלבי שיפוט.

מילות מפתח: זיהוי מגדר, כתבי יד, למידה עמוקה, רשתות נוירונים עמוקות.

מבוא

כתב יד נחשב לאחד מאופני התקשורת העתיקים ביותר בתרבות האנושית שהולך ומתפתח במהלך השנים.

אדם לומד לכתוב על ידי העתקת צורות המותאמות ומשתנות בהתאם למספר נסיבות כמו מיקום גאוגרפי, זמנים שונים וכן אופי חברתי ותרבותי.

סיווג מגדרי לפי כתב יד היא אחת הבעיות הנחקרות ביותר כיום ומעוררת עניין רב בשל מגוון התחומים הרחב שנעשה בהם שימוש בסיווג כתבי יד, החל מרפואה לצרכי מחקר על התנהגות ובריאות האדם ועד לשימושים נרחבים כמו אימות נתונים, ניתוח מסמכים היסטוריים וחקירות קרימינולוגיות.

למרות מחקרים נרחבים בנושא שנעשו במהלך השנים, בעיה זו נחשבת עדיין למאתגרת במיוחד, שכן לא ניתוחים ממוחשבים וגם לא בני אדם הציגו תוצאות מדויקות למשימה זו.

ניתן להבחין כי ברוב המקרים, טקסט בכתב יד נושא תוספת מידע על האדם שהפיק את אותו הטקסט מה שהופך את משימת ניתוח כתבי היד לאטרקטיבית במיוחד.

מחקרים פסיכולוגיים לניתוח כתבי יד אישרו כי ניתן לבצע סיווג בזכות כמה הבדלים מובהקים, בעוד שכתבי היד הנשיים נוטים להיות אחידים, מסודרים וקבועים יותר, כתבי היד של גברים נוטים להיות מחודדים, מבולגנים ומלוכסנים.

בעת ההתקדמות הטכנולוגית בניתוח תמונות ובדפוס טכניקות הסיווג, ניתוח ידני של כתב יד מוחלף במערכות אוטומטיות.

בגישה המסורתית יותר מחקרים רבים נעזרו ב-"Machine learning" לסיווג כתבי היד, בעוד שמחקרים חדשניים יותר נעזרו בכלים מתחום ה- "Deep learning".

בפרויקט זה נבצע תחילה סקירה ספרותית מקיפה תוך התמקדות בשיטות והניסויים השונים שנעשו במהלך השנים כולל הצגת התוצאות לסיווג כתבי יד, לאחר מכן נציג את שיטת העבודה ותיאור על אופי המערכת שפותחה ולבסוף נציג את תוצאות הסיווג של המערכת תוך התייחסות למאגרי הנתונים השונים שנבדקו בפרויקט זה.

המטרות העיקריות והמוטיבציה בראש ובראשונה להציג שיפור לתוצאות הקיימות שהצגנו, וכן ביצוע סיווג על מאגר נתונים בשפה העברית, ובנוסף לכל בניית מערכת אוטומטית ככלי חינמי ופתוח שישמש את כלל המשתמשים לבצע סיווג אוטומטי לתמונות כתבי יד.

הגדרת הבעיה והמטרות

במסגרת מחקרים וניסויים שנעשו במהלך השנים בנושא זיהוי מין הכותב לפי תמונת כתב לא הושגו תוצאות מספיק מדויקות הן על ידי בני אדם והן באמצעות שימוש בכלים אוטומטים, דבר המשאיר מקום לניסוים נוספים ומוטביציה לשיפור התוצאות, כמו כן המחקרים התבססו על מאגרי מידע בשפות שונות לרבות אנגלית וערבית אך עבור השפה העברית לא בוצעו מספיק מחקרים ולא קיימים מאגרי נתונים נגישים.

בנוסף, משימת הסיווג הנעשית באופן ידני על ידי בנאדם, דורשת מומחה הבקיא בניתוח המאפיינים השונים של כתב וכן מצריכה עבודה מרובה ותהליך ארוך.

למיטב ידיעתנו לא קיימת אפליקציה לזיהוי מין מכתב יד בצורה אוטומטית הנגישה לקהל הרחב.

במסגרת פרויקט זה נפתח כלי ידידותי שיהיה פתוח ונגיש לקהל הרחב, החל מחוקרים ועד למשתמשים המעוניינים לבצע ניסויים של זיהוי כתב ידם.

מטרתינו היא שהכלי בראש ובראשונה יהיה אמין ויציג אחוזי דיוק גבוהים עבור זיהוי המין, כמו כן תהליך העיבוד החל מהעלאת התמונה ועד לקבלת תוצאת הסיווג יהיה מהיר ויעיל.

כמו כן, המערכת תהיה מותאמת עבור ניתוח כתבי יד בשפה העברית.

סקירה ספרותית

1. מבוא

בעידן חדש של טכנולוגיה וכלים חדשניים, עולות תמיד בעיות שנחקרו היטב במהלך ההיסטוריה שכיום מנסים לתת להן פתרונות חדשניים וזווית הסתכלות שונה.

סיווג כתב יד היא אחת הבעיות השכיחות ולמרות זאת בוצעו דיי מעט מחקרים בנושא.

הסיווג מתפרס בעיקר לקטגוריות של זיהוי מין, גיל ולאום, אם כי זיהוי מין הוא הנפוץ והנחקר ביותר מבניהם, וגם בקטגוריה זו נכנסו לעומקה וחקרו בנוסף למין הכותב גם באיזו יד הוא השתמש בעת הכתיבה.

למעשה, זוהי נחשבת לבעיה מאתגרת במיוחד, שכן לא ניתוחים ממוחשבים בעזרת כלים של למידה מעמיקה ולא בני אדם השיגו תוצאות מדויקות למשימה זו.

משימת סיווג כתב יד משמשת כיום לחוקרים ומחקרים רבים בתחומים נרחבים החל מתחום הזיהוי הפלילי ועד לשיתופי פעולה בתחומי הרפואה, פסיכולוגיה, סוציולוגיה, ותחומים רבים נוספים שמבוססים על זיהוי מגדרי, לכן גילוי מגדר באמצעות כתב יד יכול לזרז מחקר בתחומים אחרים.

מתוך המחקרים המועטים שהיו בנושא ניתן לראות תוכנית עבודה עקבית ולרוב דומה הכוללת שני שלבים.

השלב הראשון הוא חילוץ ולמידה של מאפיינים גיאומטריים, כמו לדוגמא שיפוע, עקמומיות, מרקם וקריאות, ולאחר מכן, בשלב השני הכנסת הממצאים לתוך מסווגים ורשתות ואימון ובדיקות על מאגרי נתונים שונים.

מאגרי נתונים ברוב המחקרים היו זהים בשפות שאותן חקרו הכי הרבה כמו אנגלית וערבית, במחקרים האחרונים שבוצעו פיתחו שיטה חדשה הנקראת "זיקוק נתונים" (Data distillation).

שיטה זו נועדה על מנת להרחיב את מאגרי הנתונים באמצעות אימון מודל על נתונים עם תגיות ולאחר מכן ביצוע סיווג לנתונים ללא תגיות ואיחוד של כל הנתונים ביחד ליצירת מאגר גדול יותר.

ניתן לסקור את השיטות השונות והמחקרים שבוצעו בטבלה[11] מאגדת את כל הנתונים.

2. מחקרים וגישות

2.1 גישה פסיכולוגית

מספר גישות וטכניקות של למידת מכונה נחקרו ויושמו במהלך השנים במטרה להפיק ולקבל את תוצאות הסיווג המדויקות ביותר שניתן להשיג.

ישנם מחקרים [5] שהשתמשו אך ורק בכלים פסיכולוגיים במטרה לחלץ מאפיינים ייחודיים מתוך כתבי היד, על פי המחקר גילו שכתב יד של האדם

משתנה לאורך חייו ומושפע מגורמים כגון: גיל, מצב נפשי בעת הכתיבה ומידת הריכוז.

כדי לבדוק את הדגימות ואת המאפיינים שקיימים בכל דגימה הם השתמשו בכלים סטטיסטים כדי לבחון את הבעיה.

לאחר קבלת התוצאות ובחינת הממצאים הם הגיעו למסקנה כי יש הבדלים מהותיים בין כתב יד של גבר לבין כתב יד של אישה, ויש שוני במאפייני הכתב, לכן ניתן לבדוק עבור דגימות כתבי יד האם הדגימה נכתבה ע"י גבר או ע"י אישה לפי מאפיינים אלו.

2.2. מודלים ולמידה עמוקה

2.2.1. SVM - support vector machines

ניתן לראות כי מודל הסיווג הפופולארי מתוך המחקרים שנבדקו היה SVM, על פי [2], נבדקו שני מערכות סיווג שונות, הראשונה מבוססת על SVM (support vector machines) והשנייה מבוססת על GMM (Gaussian Mixtures Models).

החידוש שנעשה בנוסף לסיווג מין הכותב הוא גם באיזו יד השתמש הכותב, ולכן שני מערכות הסיווג אומנו לפתור את שתי הבעיות על אותו מאגר כתבי יד הנקרא IAM-OnDB שכלל כתבי יד של 200 משתתפים ולשם השוואה נבדק סיווג ע"י בני אדם.

לפי התוצאות שקיבלו, GMM זכה לאחוזי הצלחה גבוהים יותר, 67% לזיהוי מין ו- 84% לזיהוי יד הכותב, בעוד ש SVM קיבל 61% אחוזי הצלחה עבור 2 בעיות הסיווג.

תוצאות הניסוי שערכו לזיהוי הבעיות ע"י בני אדם היו נמוכות יותר ועמדו על 57% לזיהוי מין הכותב ו 62% עבור זיהוי יד הכותב.

2.2.2. ANN - artificial neural networks

מחקר נוסף [3] שכלל שימוש במודל הסיווג מסוג SVM, השתמש בנוסף במסווג מסוג ANN (artificial neural networks), ושניהם אומנו על שני מאגרי נתונים שונים ומאוד פופולאריים בקרב החוקרים, הראשון QUWI - Qatar University Writer Identification, והשני MSHD - Multiscript Handwritten Database.

מסווג הANN הכיל רשת בת שלוש שכבות: שכבת הקלט הכילה את מספר הנוירונים לפי כמות התכונות שזיהו בכל כתב, שכבת הפלט הכילה שני נוירונים לפי זיהוי המין זכר ונקבה, ושכבת הביניים הכיל מספר נוירונים כפונקציה של הממדיות של וקטור תכונת הקלט.

כל נוירון הכיל פונקציות העברה מסוג sigmoid והרשתות אומנו באמצעות "back propagation algorithm".

נערכו מספר ניסויים עבור כל מאגר מידע ותוצאות הסיווג עמדו על 68.75% עבור מאגר הנתונים QUWI ו73.02% עבור MSHD.

2.2.3. שילוב תכונות פסיכולוגיות ולמידה עמוקה

במחקר חדשני יותר [8] הוצגה שיטה חדשה שנערכה על מאגר הנתונים הידוע QUWI המשלבת את מודל ה- SVM הקובע את הסיווג על פי המקסימום בין שני המאפיינים COLD ו - Hinge.

התכונה COLD מחלצת צורות ייחודיות של רכיבי טקסט בכתב יד על ידי ניתוח הקשר בין נקודות דומיננטיות כמו תכונות ישרות ומכוונות זווית ועקמומיות על קווי מתאר של רכיבי טקסט בכתב יד.

התכונה Hinge – מבוססת גם כן על היבט פסיכולוגי באפיון צורת הכתב.

תוצאות המחקר מבוססות על שלושה ניסויים שבוצעו כאשר בכל ניסוי הוכנסו פרמטרים שונים עבור התכונות.

בניסוי הראשון הגיעו לממוצע של 73.60% אחוז בזיהוי, בניסוי השני ממוצע של 74.20% ובשלישי 64.40%, השוני בתוצאות נבע משיטות שונות לבדיקה שנערכו כאשר בראשון השתמשו בתכונות שונות בשימוש עם המסווג, בניסוי השני השתמשו בשיטת Script-dependent ובשלישי script-independent.

2.2.4. אלגוריתם יער אקראי ו- KDA

בשונה מעט מהמחקרים שהוצגו עד כה ישנם מחקרים ששילבו כלים ומסווגים אחרים.

על פי [1] הציעו מספר מאפיינים גאומטריים, לדוגמא: עקמומיות וכיוון הכתב, לאפיון כתב היד שישולבו במסווגים וכלים כגון אלגוריתם יער אקראי ו- KDA.

התוצאות שהתקבלו מתוך הניסוי שנערך על מאגר הנתונים שהוצג לעיל מגיעים לסיווג של 74.05% למין כותב, 55.76% עבור הגיל ו- 53.66% עבור הלאום כאשר כולם כתבו את אותו הטקסט, ועבור טקסט שונה התקבלו תוצאות מעט יותר נמוכות 73.59% עבור מין הכותב, 60.62% לגיל ו-47.98% עבור הלאום.

המסקנות שהתקבלו מתוך התוצאות הן שעדיף יהיה להשתמש באלגוריתם היער האקראי עבור סיווג של גיל ולאום ואילו עבור המין KDA הציג תוצאות טובות יותר.

מסקנה נוספת שהתגלתה היא שכתבי יד שהופקו על ידי אותו כותב מניבים תוצאות מעט טובות יותר לסיווג מגדר אך לא עבור סיווג טווח הגילאים או הלאום.

2.2.5. עצי החלטה

ניתן לראות כי גם ע"פ [6] נעשה המחקר בעזרת שימוש בגרפולוגיה מצד אחד וכלים מעולם מדעי המחשב מצד אחר.

בשלב הראשון של הניסוי (האם כותב יכול לזהות את כתב ידו מתוך המאגר) קיבלו כי 47% מהמשתתפים יכלו לזהות את כתב ידם.

לחלק השני של הניסוי(זיהוי מין הכותב) החוקרים יצרו סט של חוקים הכוללים 133 מאפיינים ובנו 2 עצי החלטה כדי להכריע את הסוגייה. עץ ההחלטה הראשון נבנה בעזרת שימוש באלגוריתם j48, 70% מהנתונים שימשו לאימון בעוד שה 30% הנותרים שימשו לבדיקה. בעזרת שימוש בעץ זה קיבלו תוצאה של 70% בסיווג נכון של מין הכותב, בנוסף 8 מתוך 133 המאפיינים זוהו כמספקים עבור הכרעה.

עץ ההחלטה השני נבנה באמצעות אלגוריתם id3,השתמשו 80% מהמידע לצורך אימון ו 20% הנותרים לצורך בדיקה, אחוזי ההצלחה היו גבוהים והגיעו ל- 93%.

החוקרים הגיעו למסקנה כי יש מאפיינים המבדילים בכתב בין המינים ושניתן לבצע משימה זאת בעזרת שימוש במדעי המחשב וכריית מידע.

2.2.6. CNN - convolutional neural networks

מודל הסיווג שהציג עד עתה את תוצאות הסיווג המדויקות ביותר היה מודל ה-CNN (convolutional neural networks).

גם במחקר הנ"ל [9] ביצעו את הניסוי בשני חלקים שכללו גם סיווג של מין הכותב על פי כתב ידו וגם באיזו יד השתמש על מנת לכתוב על שני מאגרי נתונים של כתבי יד ציבוריים, הראשון IAM המכיל טקסטים באנגלית, והשני KHATT המכיל טקסטים בערבית.

המודל נבנה מ6 שכבות מאומנות ה-2 הראשונות מסוג convolutional, וה-2 האחרונות מסוג dense, הרשת קיבלה בשלב הראשוני קלט תמונה בגודל 30X100 .

עבור מאגר הנתונים IAM, התקבלו תוצאות סיווג של 80.72% עבור מין ו- 90.70% עבור יד הכותב.

עבור מאגר הנתונים KHATT, התקבלו תוצאות סיווג של 68.90% עבור מין ו – 70.91% עבור יד הכותב.

המחקר הישראלי [10] גם כן ביצע שימוש במסווג הCNN, לטענתם, אחוזי ההצלחה שהציגו דומים לשיטות טכנולוגיות אחרות שנבדקו בנושא, ואחוזי ההצלחה שהתקבלו לעומת בדיקה של בני אדם היו גבוהים יותר.

האימון נעשה על מאגר נתונים פרטי שהוכן במיוחד, הם חילקו ל 405 משתתפים טופס, המשתתפים מילאו פרטים אישיים שונים ולכתוב טקסט מסוים. מכיוון שיש במאגר פרטים נוספים בנוסף למין הכותב ניתן להשתמש במאגר למטרות שונות.

על מאגר המידע הם ביצעו עיבוד מקדים שבו הם לקחו את החלק של הטקסט, העבירו אותו לגוון אפור וחילוקו את הטקסט למספר חלקים רנדומליים בצורה של מלבן ושל מרובע (לשורה ולמילה).

לאחר העיבוד המקדים הם השתמשו בנתונים כדי לאמן את מודל ה CNN שלהם.

את בעיית הזיהוי הם חילקו לשלושה חלקים:

Intra – language : האימון והטסט נעשים על שפה אחת – במצב זה אחוזי הדיוק היו על שתי השפות (עברית ואנגלית) היו בממוצע 73% - 75%.

Inter – language : האימון נעשה על שפה אחת והטסט נעשה על השפה שנייה – במצב זה אחוזי הדיוק היו 75% וגם 59%.

Mixed language classification – משלבים את כתבי היד באנגלית וכתבי היד בעברית למאגר אחד ובכך מגדילים את המאגר. כך שהאימון והטסט נעשה על שתי המאגרים – במצב זה אחוזי הדיוק היו 77% .

כדי להשוות את התוצאות לתוצאות זיהוי של בני אדם הם בנו תכנית בשביל 300 משתתפים שהתבקשו להכריע לגבי 15 כתבי טקסט באנגלית ו 15 כתבי טקסט בעברית האם נכתבו על ידי גבר או אישה. תוצאות הדיוק היו בערך 65% .

על מאגר המידע הם ביצעו עיבוד מקדים שבו הם לקחו את החלק של הטקסט, העבירו אותו לגוון אפור וחילוקו את הטקסט למספר חלקים רנדומליים בצורה של מלבן ושל מרובע (לשורה ולמילה).

לאחר העיבוד המקדים הם השתמשו בנתונים כדי לאמן את מודל ה CNN שלהם.

את בעיית הזיהוי הם חילקו לשלושה חלקים:

1. Intra – language : האימון והטסט נעשים על שפה אחת – במצב זה אחוזי הדיוק היו על שתי השפות (עברית ואנגלית) היו בממוצע 73% - 75%.
2. Inter – language : האימון נעשה על שפה אחת והטסט נעשה על השפה שנייה – במצב זה אחוזי הדיוק היו 75% וגם 59%.
3. Mixed language classification – משלבים את כתבי היד באנגלית וכתבי היד בעברית למאגר אחד ובכך מגדילים את המאגר. כך שהאימון והטסט נעשה על שתי המאגרים – במצב זה אחוזי הדיוק היו 77% .

כדי להשוות את התוצאות לתוצאות זיהוי של בני אדם הם בנו תכנית בשביל 300 משתתפים שהתבקשו להכריע לגבי 15 כתבי טקסט באנגלית ו 15 כתבי טקסט בעברית האם נכתבו על ידי גבר או אישה. תוצאות הדיוק היו בערך 65% .

2.2.7. טכניקת מורה-תלמיד

ללמידה מעמיקה הישגים יפים בתחום עיבוד וזיהוי תמונה בשנים האחרונות, עם זאת רוב המודלים מאומנים לפי גישת supervised learning , כלומר הלמידה נעשית על תמונות בעלות תגיות ולכן נדרש מאגר תמונות גדול כדי שהמודל ילמד היטב.

לפיכך, אנו מגבילים את עצמנו לשימוש רק בתמונות בעלות תגיות, למרות שישנן כמויות גדולות הרבה יותר של תמונות ללא תגיות הניתנות לשימוש.

כדי לפתור בעיה זאת הציעו [7] שיטה של לימוד עצמי עבור המודל תוך שימוש בתמונות בעלות תגיות ותמונות ללא תגיות.

לשיטה 3 שלבים עיקריים:

1.אימון מודל "מורה" – מאמנים מודל על התמונות בעלות התגיות.

2. משתמשים במודל המורה כדי לסווג את התמונות שאין להן תגיות.

3. מאמנים מודל חדש – מודל "תלמיד" שמאומן על התמונות המסווגות וגם על התמונות שהיו לא מסווגות ומודל המורה סיווג אותם.

את התהליך מבצעים במספר איטרציות כך שבכל איטרציה מודל הסטודנט הופך להיות מודל המורה החדש ועכשיו תפקידו לסווג מחדש את התמונות הלא מסווגות. בכך הם מפחיתים את ה- combined cross entropy loss על שני סוגי התמונות. בניסוים שביצעו קיבלו את התוצאה הטובה לאחר שלוש איטרציות.

בניסויים שביצעו הם ראו שכדי שהתהליך יעבוד היטב מוטב כי בתהליך הלמידה של המורה לא כדאי שיהיה רעש כדי שהתוצאות יהיו מדויקות ככל האפשר, אך בתהליך הלמידה של מודל הסטודנט כדאי שיהיה רעש כדי שהוא יצטרך ללמוד טוב יותר על מקרים קשים יותר לסיווג כדי להתעלות על התגיות שמודל המורה הפיק.

בשימוש בשיטה זאת החוקרים השיגו אחוזי דיוק של 88.4 שהיו גבוהים יותר ב 2-3% מאחוזי דיוק בשימוש רשת זאת לסיווג תמונות ללא השיטה.

2.2.8. "זיקוק נתונים" (Data Distillation)

מחקר נוסף [4] הציע שיטה דומה לפתרון בעיית התגיות והגדלת מאגר הנתונים כדי לקבל תוצאות טובות יותר, השיטה נקראת "זיקוק נתונים" (Data Distillation).

הרעיון הוא לאמן מודל באמצעות כמות גדולה מאוד של נתונים, לבצע סיווג ואז לאמן שוב את המודל באמצעות התגיות החדשות שנוצרו, עם זאת ישנה בעייתיות של מידע חסר חשיבות שנוצר מאימון מודל על התחזיות שהוא עצמו מייצר.

הטיפול בבעיה זו לשיטת החוקרים היא ע"י הרכבת תוצאות מסיווגים שונים של מודל יחיד לנתונים ללא תגית, טרנספורמציות כאלה הוכיחו את יעילותם בשיפור אחוז הדיוק של המודל.

על מנת לבחון את שיטת זיקוק הנתונים, השתמשו במשימה לזיהוי נקודות מפתח אנושיות(human keypoint detection task) על מאגר נתונים הנקרא COCO [4], המכיל דגימות ותמונות מהעולם האמיתי.

האימון התבצע על מודל הנקרא Mask R-CNN [4] ובנוסף שימוש בשיטת זיקוק הנתונים על התגיות המקוריות הידועות במאגר הנתונים ועל כמות גדולה נוספת של נתונים ללא תגית.

חלוקת הנתונים הייתה של כ-115 אלף תמונות עם תגיות שנקראו 115-CO, ו120 אלף תמונות ללא תגיות שנקראו UN-120, ובנוסף כ180 אלף תמונות מתוך מאגר בשם Sports-1M עבור זיקוק הנתונים שנקרא s1m-180.

הניסויים בוצעו בחלוקה ל-3 קבוצות:

1. מאגר קטן שהכיל 35 אלף תמונות עם תגית ועוד 80 אלף תמונות ללא תגית.
2. מאגר גדול שהכיל 115 אלף תמונות עם תגית(CO-115), ו120 אלף תמונות ללא תגית(UN-120).
3. מאגר גדול שהכיל 115 אלף תמונות עם תגית(CO-115), ו180 אלף תמונות ללא תגית(s1m-180).

התוצאות הראו שיפור קל באחוז הדיוק של הסיווגים בין מאגר עם תגיות ידועות מראש שהציג 73.6 אחוז דיוק למול מאגר שביצעו עליו זיקוק נתונים שהציג 75 אחוז דיוק, ומכאן עלתה המסקנה כי ניתן לייעל ולעקוף את התוצאות של למידה מפוקחת באמצעות השיטה החדשה שהדגימו במאמר זה שהציג תוצאות טובות יותר.

3. סיכום

נושא סיווג כתב יד הוא חשוב מאוד למחקרים וניסויים במגוון רחב של תחומים.

הנושא אומנם נחקר במידה יחסית מועטה אך ישנם מחקרים שהניבו תוצאות מדויקות בעזרת שימוש בכלים שונים מתחום מדעי המחשב.

בנוסף למוטיבציה להשגת התוצאות המדויקות יותר, חוקרים מנסים גם לבחון ניסויים נוספים במקביל לסיווג מין כמו סיווג היד בה השתמש הכותב בעת כתיבת הטקסט.

אין ספק כי הנושא ימשיך להיות מרכזי בתחום כריית המידע ולימוד המכונה, חוקרים ימשיכו לנסות להרחיב גבולות ולנסות להשיג תוצאות מדויקות ככל הניתן באמצעות שילוב של כלים הן מתחומים של פסיכולוגיה והן מתחום המסווגים והכלים של מדעי המחשב.

במחקר שלנו נעבוד על מאגר נתונים פרטי שהוכן מבעוד מועד במטרה להרחיב את מאגרי כתבי היד בשפה העברית.

המודל שלנו יבוסס על רשת CNN כשלנגד ענינו הרצון לייעל ולשפר ככל הניתן את אחוז הדיוק בסיווג מין על פי כתב יד.

4. השוואה בין המתודולוגיות שנסקרו

מאמר משנת	מטרת המאמר	מאגר נתונים בו השתמשו	תוכנות בהן השתמשו	מסווג	תוצאות
2014	סיווג אוטומטי של גיל, מין ולאום על סמך כתב יד [1]	QUWI המכיל 1017 טקסטים באנגלית וערבית	עקמומיות, כיוון, chain code	Random forest KDA	KDA – 73.6% RF – 74.8%
2015	סיווג מין הכותב על סמך כתב יד [3]	QUWI המכיל 475 טקסטים באנגלית וערבית MSHD שהכיל 87 טקסטים בצרפתית וערבית	נטיית הכתב, עקמומיות, מעוגלות, סדר, קריאות ומרקם	ANN SVM	QUWI SVM 68.75% ANN- 67.50%/ MSHD SVM- 73.02% MSHD- 69.44%
2018	ניתוח וסיווג מין הכותב ובאיזה יד הוא השתמש לכתיבה [9]	IAM המכיל 657 טקסטים באנגלית KHATT המכיל כ1000 טקסטים בערבית	X	CNN	IAM מין – 80.72% יד – 90.70% KHATT מין – 68.90% יד – 70.91%
2020	סיווג מין כותב על סמך כתב יד באמצעות תכונות חדשות [8]	QUWI שהכיל 750 טקסטים בערבית	1. COLD -Cloud of Line Distribution 2.Hinge	SVM	שלושה ניסויים: 1. 73.60% 2. 74.20% 3. 64.40%
2018	סיווג מין כותב על סמך כתבי יד באמצעות רשתות נוירוניות [10]	מאגר כתבי יד שהוכן במיוחד המכיל 405 טקסטים	X	CNN	התוצאות חולקו לפי 3 קטגוריות: 1. Intra – language 75%-73% 2. Inter – language 75% ו59% 3. Mixed – 77%
2007	ניתוח וסיווג מין הכותב והיד בה השתמש לכתיבה [2]	IAM-OnDB הכולל 200 משתתפים	X	GMM SVM	GMM מין – 67% יד – 84% SVM מין – 61% יד – 61%
2017	גילוי מאפיינים בכתב יד לזיהוי המין לצורך זיהוי פלילי [5]	מאגר כתבי יד פרטי המכיל טפסים שחולקו בין 130 משתתפים	27 מאפיינים יחודיים הכוללים בין היתר נטיית הכתב, מעוגלות, סדר וקריאות	X	X
2017	זיהוי מין הכותב באמצעות שימוש בגרפולוגיה וכלים ממדעי המחשב [6]	מאגר כתבי יד פרטי בשפה התורכית המכיל טפסים שחולקו בין 80 משתתפים	מאפיינים כגון: לחץ העט, לכסון, ממדי הכתב	עץ החלטה שנבנה באמצעות 2 אלגוריתמים: J48 ID3	עץ החלטה j48 – 70% עץ החלטה id3 – 93%
2020	סיווג תגיות לתמונות בעזרת שיטת תלמיד-מורה [7]	ImageNet המכיל כ-14 מיליון תמונות עם וללא תגיות	Semi-supervised learning. Student-teacher model.	EfficientNet(CNN)	88.4%
2017	סיווג תגיות באמצעות שיטת זיקוק נתונים [4]	COCO המכיל כ310 אלף תמונות שמתוכן כ220 אלף עם תגיות	Data Distillation. Student-teacher model.	Mask R-CNN	75%

שיטת העבודה

1. תיאור מאגר הנתונים - HHD_gender

מאגר הנתונים כולל 819 טפסים, כאשר בכל טופס המשתמש התבקש להזין פרטים אישיים כמו עיר מגורים, מין וגיל.

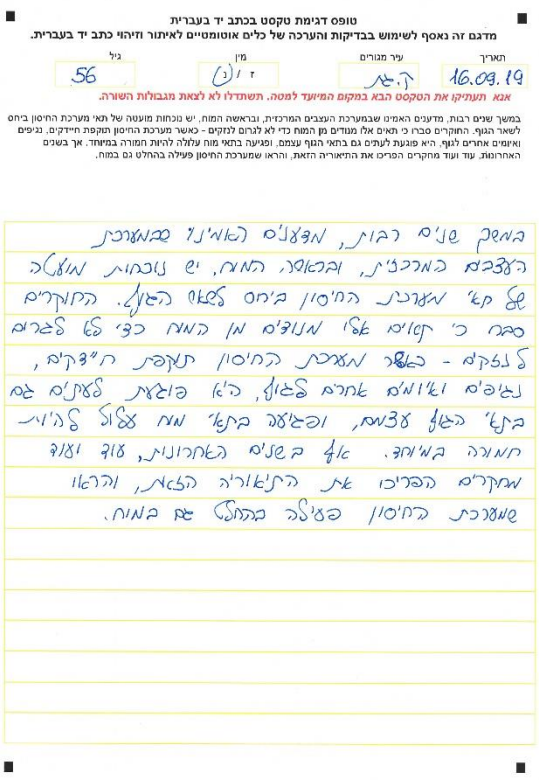
בכל טופס ישנה פסקה הלקוחה מבין 50 סוגים שונים של פסקאות שאותה המשתמש התבקש לכתוב בכתב ידו במסגרת המסומנת עם קווים צהובים למילוי.

במאגר ישנם 358 טפסים שנכתבו על ידי משתמשים ממין זכר, ו461 טפסים שנכתבו על יד משתמשים ממין נקבה.

טווח הגילאים של המשתמשים שהתבקשו למלא את הטופס נא בין 10-60.

כל טופס נסרק ברזולוציה של 6843 X 4816, ובפורמט JPG.

הטפסים נשמרו במאגר מקוון בGoogle Drive.



טופס דגימת טקסט בכתב יד בעברית

מדגם זה נאסף לשימוש בבדיקות והערכה של כלים אוטומטיים לאיתור וזיהוי כתב יד בעברית.

תאריך	עיר מגורים	מין	גיל
16.09.19	קיבוץ	זכר	56

אנא תעתיקו את הטקסט הבא במקום המיועד למטה. תשתדלו לא לצאת מגבולות השורה.

במשך שנים רבות, מדענים האמינו שבמערכת העצבים המרכזית, ובראשה המוח, יש נוכחות מועטה של תא' מערכת החיסון בי"חס לשאר הגוף. החוקרים סברו כי תאים אלו מגודים מן המוח כדי לא לגרום לנזקים - כאשר מערכת החיסון הוקפה חיידקים, נגיפים ואיומים אחרים לגוף, היא פוגעת לעתים גם בתאי הגוף עצמם, ופגיעה בתאי מוח עלולה להיות חמורה במיוחד. אך בשנים האחרונות, עוד ועוד מחקרים הפריכו את התיאוריה הזאת, והראו שמערכת החיסון פעילה בהחלט גם במוח.

במשך שנים רבות, מדענים האמינו שבמערכת
העצבים המרכזית, ובראשה המוח, יש נוכחות מועטה
של תא' מערכת החיסון ביחס לשאר הגוף. החוקרים
סברו כי תאים אלו מגודים מן המוח כדי לא לגרום
לנזקים - כאשר מערכת החיסון תוקפת חיידקים,
נגיפים ואיומים אחרים לגוף, היא פוגעת לעתים גם
בתאי הגוף עצמם, ופגיעה בתאי מוח עלולה להיות
חמורה במיוחד. אך בשנים האחרונות, עוד ועוד
מחקרים הפריכו את התיאוריה הזאת, והראו
שמערכת החיסון פעילה בהחלט גם במוח.

1

2. עיבוד מקדים

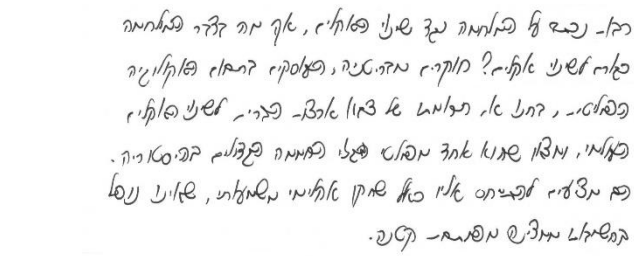
בשלב הראשון בוצע יישור לטופס באמצעות חישוב סטיות שאולי התקבלו במהלך הסריקה של הטפסים.

לאחר מכן, כדי לבצע מיון של כל הטפסים לפי מין הכותב, התבצע אלגוריתם המאתר את התיבה שבו התבקש המשתמש לסמן את מינו.

לאחר איתור התיבה, האלגוריתם מחלק את התיבה לשני חלקים ומסווג לפי זיהוי הסימון של המשתמש כאשר סימון בצד ימין יסווג למין זכר, וסימון בצד שמאל יסווג למין נקבה.

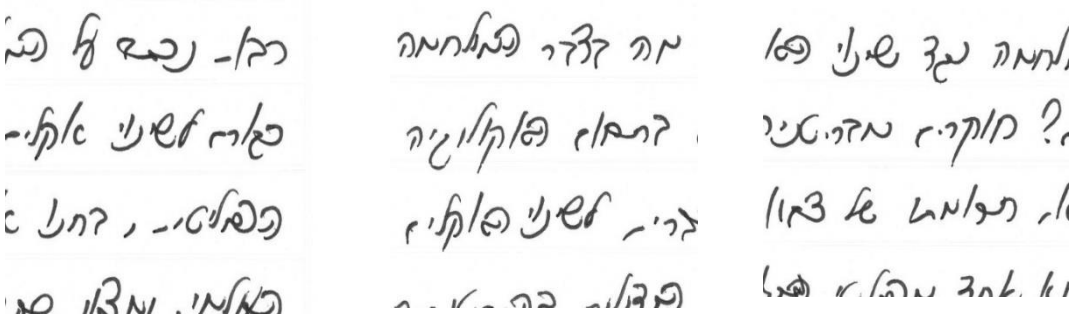
כאשר מין הכותב זוהה, מתבצע חילוץ אזור הטקסט מהטופס, כך שאזור זה יכיל את כתב היד בלבד, ובנוסף על ידי שימוש בטכניקות של עיבוד תמונה, מתבצעת הסרה של קווי השורות הצהובים והמרת גוון התמונה לגווני אפור.

בעת סיום העיבוד המקדים לכל טופס מן המאגר, מתבצעת חלוקה לסט של Train(80%), Test(10%), Validation(10).



3. חלוקה ל- Patches

בשלב זה כאשר התמונות מוכנות לעבודה, מתבצע תהליך חלוקת התמונות ל- Patches.
תהליך זה מתבסס על חלוקת תמונה דיגיטלית למקטעים כאוספים של פיקסלים הנמצאים זה ליד זה בתמונה.
מטרת התהליך היא לפשט או לשנות את הייצוג של התמונה לאוסף של אובייקטים בעלי משמעות שניתן להתייחס אליהם לאחר מכן על ידי אלגוריתמים אחרים לראייה ממוחשבת או לעיבוד תמונה ,כגון מציאת גבולות או קווים.
התוצאה של התהליך היא אוסף של מקטעים זרים המכסים את התמונה כולה.
שיטת החלוקה ל-patches מתבצעת על ידי "חלון הזזה", גודל החלון הינו 400X400, ובאמצעות תזוזה מהקצה השמאלי של התמונה, התמונה מחולקת ל-patches כאשר במקביל מתבצעת בדיקה שיש מספיק מידע – מספיק פיקסלים בכל patch.
לכל טופס מתקבלת כמות שונה של patches, בתלות כמות הכתב בכל טופס.
תמונות אלה נשמרו במאגר נוסף לפי החלוקה שהוגדרה בשלב העיבוד המקדים(Train, Test, Validation).



4. אימון המודלים

כדי לסווג כל patch, מתבצע אימון של מספר מודלים לזיהוי תמונה מאומנים מראש.
מודלים אלו אומנו לזהות אלף מחלקות שונות של תמונות לפי מאגר התמונות ImageNet.
למודלים אלו מבצעים Fine-tuning, כדי להשתמש במאפיינים שהמודלים למדו בעבר על מנת ללמד אותם לזהות מחלקות שלא הכירו לפני כן.
Fine-tuning הינו תהליך הכולל מספר שלבים:

- הסרת שכבת ה-Fully connected האחרונה במודל.
- הוספת שכבה חדשה בעלת שתי יחידות.
- הקפאת שאר השכבות במודל.
- אימון המודל על מאגר חדש כדי לזהות מחלקות חדשות.

המודלים שנבחנו במסגרת הפרויקט הם:

- Xception
- EfficientNet B0
- NasNet
- Vgg16
- Inception_V3
- DenseNet121
- DenseNet169
- Vgg19
- Baseline – מודל CNN בסיסי הבנוי מ-11 שכבות שאינו pre-trained כמו שאר המודלים, כלומר לא אומן קודם לכן על שום מאגר נתונים.

5. הערכת התוצאות

לאחר אימון המודלים, מתבצעת תהליך הסיווג של המודלים על קבוצת התמונות שהוגדרו כמאגר הTest.

לכל תמונה במאגר זה, מתבצעת חלוקה ל- Patches רבים ככל האפשר לפי אלגוריתם "חלון ההזזה", כל מודל בתורו מסווג כל patch באופן נפרד למחלקה שזיהה, זכר או נקבה, לפי תוצאות אילו מתבצע סיווג סופי למסמך הטקסט לפי שיטת הרוב, שבה המחלקה שקיבלה מספר הצבעות גבוה יותר, היא המחלקה שתקבע כסיווג הסופי.

לאחר מכן מתבצעת בדיקה האם המחלקה שהמודל סיווג היא באמת המחלקה הנכונה.

מתבצעת ספירה של כמות הפעמים בהם המודל סיווג טקסט באופן נכון, אחוז הדיוק(Accuracy) שנמדד לכל מודל הינו כמות ההחלטות הנכונות מסך התמונות במאגר.

אחוז דיוק(Accuracy) של מערכת מדידה נקבעת על פי מידת הקרבה של המדידות של כמות מסוימת, לערך הממשי האמיתי של אותה כמות .

$$\text{דיוק} = \frac{\text{מספר התוצאות הנכונות החיוביות}}{\text{מספר התוצאות החיוביות הנכונות+ החיוביות הלא נכונות}}$$

מסמך ייזום

1. תקציר מנהלים

1.1 יעדים

סיווג כתב יד היא אחת הבעיות השכיחות ולמרות זאת בוצעו דיי מעט מחקרים בנושא.
הסיווג מתפרס בעיקר לקטגוריות של זיהוי מין, גיל ולאום.
זיהוי מין לפי כתב יד הוא הנפוץ והנחקר ביותר מבניהם, וגם בקטגוריה זו נכנסו לעומקה וחקרו בנוסף למין הכותב גם באיזו יד הוא השתמש בעת הכתיבה.
למעשה, זוהי נחשבת לבעיה מאתגרת במיוחד, שכן לא ניתוחים ממוחשבים בעזרת כלים של למידה מעמיקה ולא בני אדם השיגו תוצאות מדויקות למשימה זו.
משימת סיווג כתב יד משמשת כיום לחוקרים ומחקרים רבים בתחומים נרחבים החל מתחום הזיהוי הפלילי ועד לשיתופי פעולה בתחומי הרפואה, פסיכולוגיה, סוציולוגיה, ותחומים רבים נוספים שמבוססים על זיהוי מגדרי, לכן גילוי מגדר באמצעות כתב יד יכול לזרז מחקר בתחומים אחרים.
מכאן עולה היעד המרכזי של פרויקט זה, לפתח כלי מדויק ככל שניתן על מנת לתת פתרון יעיל וחדשני לנושא סיווג מין לפי כתבי יד.

1.2 יישום

הפתרון שאנו מציעים יעבוד לפי מספר שלבים, בשלב הראשון איסוף נתונים הכוללים תמונות של כתבי יד ועיבודן, בשלב השני אימון מודלים ידועים בתחום עיבוד התמונה על מאגר כתבי היד שלנו, בשלב השלישי הערכת התוצאות, בחירת שלושה מודלים בעלי אחוז הדיוק הגבוה ביותר, ובשלב האחרון שילוב המודלים וסיווג תמונה לפי שיטת "הרוב".

1.3 טכנולוגיה ותשתית

מספר כלים וסביבות עבודה שימשו אותנו לפיתוח המערכת, סביבת העבודה המרכזית הייתה Pychram עם שפת תכנות Python.

Google drive - אחסון נתונים של כתבי היד.

Google colab – נעזרנו בשרתים של גוגל על מנת להריץ מספר סקריפטים של אימון מודלים במקביל, כמו כן מנועי ההרצה(CPU, GPU) עזרו לנו להגיע לתוצאות מהירות יותר מאשר הרצה במחשבים הרגילים.

תשתית הפיתוח הייתה בשני מחשבים ניידים עם מערכת הפעלה Windows 10.

1.4 מימוש

בניית מערכת עבור המשתמש בו יוכל להעלות תמונת כתב יד ולקבל סיווג לפי מין, המערכת תטען את התמונה ותציג בפני המשתמש את התמונה שנטענה, במקרה של שגיאה המערכת תתריע בהתאם.

לאחר טעינת התמונה יוצגו למשתמש מספר הגדרות כמו בחירת המודלים שברצונו להשתמש על מנת לסווג, בחירת האפשרות לסיווג ע"י שילוב בין שלושה מודלים וסיווג באמצעות שיטת "החלטת הרוב".

לאחר ששלב הסיווג הסתיים בהצלחה, יוצגו למשתמש תוצאות מפורטות והסבר על התהליך שבוצע.

2. מנהלה

2.1 צוות הפרויקט:

מנחות הפרויקט :

- ד"ר אירינה רבייב
- ד"ר מרינה ליטבק

סטודנטים:

- אורי טביבי
- שון אסולין

2.2 תוכנית עבודה – פירוט תוכנית עבודה שנתית(לפי חלוקה לסמסטרים).

סמסטר א':

- 7.11.20 – הגשת שלד לסקר ספרות.
- 7.12.20 – הגשה ראשונית לקראת וועדה מלווה הכולל סקר ספרות, סקר שוק, מסמך ייזום, דרישות ועיצוב ותיאור הסביבה ופיתוח שנעשה עד כה.
- 16.12.20 – מפגש וועדה מלווה 1.
- 15.1.21 – הגשה סופית של התוצרים שנעשו עד כה.

סמסטר ב':

- 4.4.21 – מפגש וועדה מלווה 2.
- 13.5.21 – הצגת תקציר פרויקט בעברית ואנגלית לאישור המנחה.

- 18.5.21 – הצגת מצגת פרויקט מסכמת לוועדה מלווה.
- 25.5.21 – הצגת פוסטר ראשוני לאישור המנחה.
- 1.6.21 – הצגת פוסטר פרויקט סופי.
- 10.6.21 – הגשת דו"ח מסכם והשתתפות בכנס פרויקטים מחלקתי.

2.3 שיטת העבודה:

- פגישות עבודה דו-שבועיות עם מנחות הפרויקט.
- פגישות במסגרת קורס סמינר פרויקט בהתאם לסילבוס.
- פגישות יומיות לצורך עבודה ועדכונים בין הסטודנטים.

3. יעדים

3.1 לקוח/מומחי יישום

3.1.1 לקוח עיקרי/משתמש – משימת סיווג כתב יד משמשת כיום לחוקרים ומחקרים רבים בתחומים נרחבים החל מתחום הזיהוי הפלילי ועד לשיתופי פעולה בתחומי הרפואה, פסיכולוגיה, סוציולוגיה, ותחומים רבים נוספים שמבוססים על זיהוי מגדרי, לכן גילוי מגדר באמצעות כתב יד יכול לזרז מחקר בתחומים אחרים.

3.1.2 מומחי היישום

- אורי טביבי.
- שון אסולין.

3.2 מטרות ויעדים

3.2.1 מטרה כללית – יצירת כלי לסיווג מין לפי תמונות כתבי יד.

3.2.2 מטרה מעשית – אימון מודלים והערכתם על מנת לקבל אחוזי דיוק גבוהים ככל הניתן בסיווג תמונות כתבי יד.

3.2.3 מטרות לעתיד – המשך מחקר בנושא על מנת לייצר כלי מדויק ככל הניתן לסיווג כתבי יד.

3.3 בעיות

3.3.1 תיאור הבעיה במצב כיום

סיווג כתב יד היא אחת הבעיות השכיחות ביותר בתחום הלמידה המעמיקה ולמרות זאת בוצעו דיי מעט מחקרים בנושא.
למעשה, זוהי נחשבת לבעיה מאתגרת במיוחד, שכן לא ניתוחים ממוחשבים בעזרת כלים של למידה מעמיקה ולא בני אדם השיגו תוצאות מדויקות למשימה זו.
בעיה זו מתוארת בספרות כבעיה נרחבת במספר תחומים מרכזיים כמו רפואה, פסיכולוגיה, סוציולוגיה, ותחומים רבים נוספים שמבוססים על זיהוי מגדרי, לכן גילוי מגדר באמצעות כתב יד יכול לזרז מחקר בתחומים אחרים.

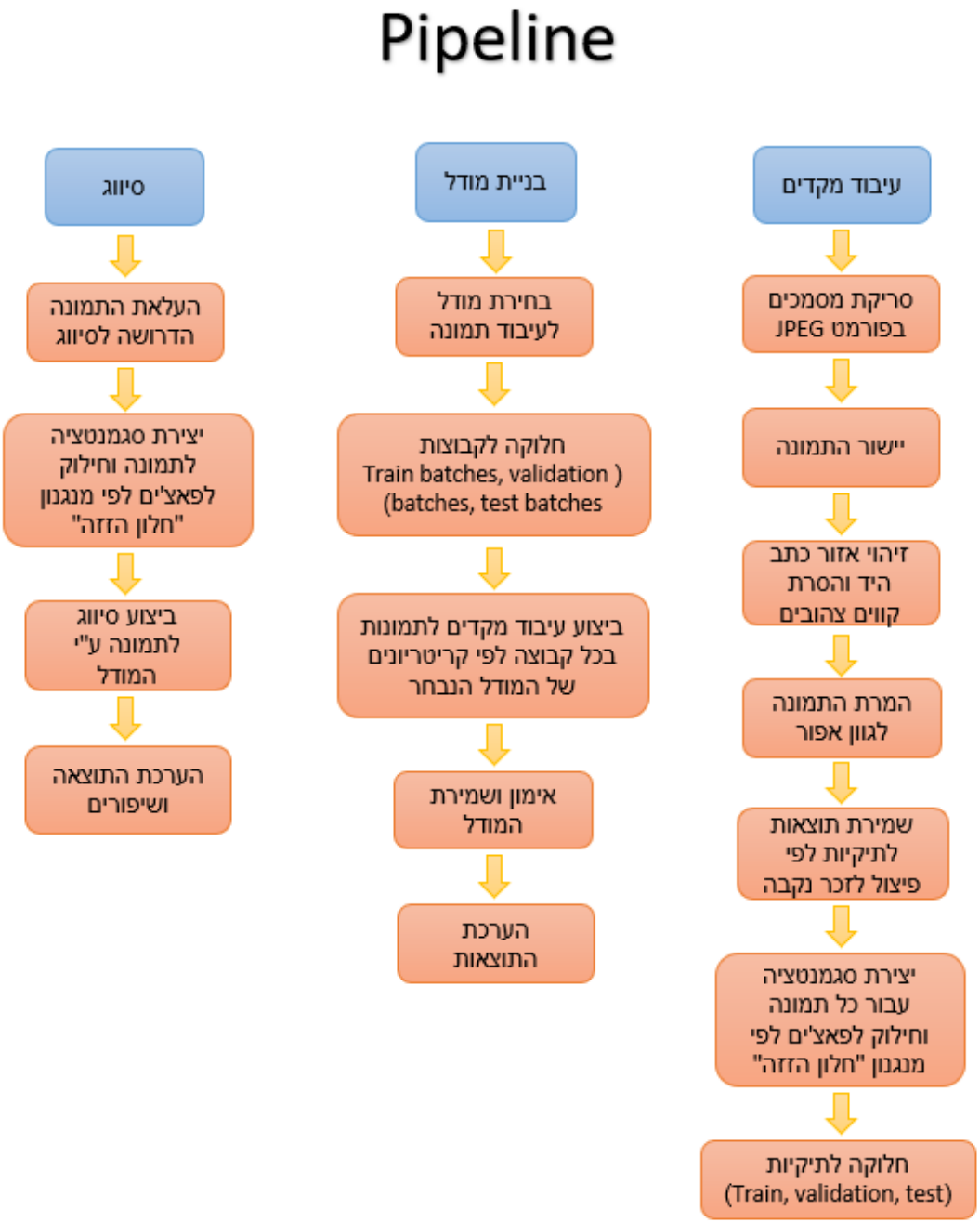
3.3.2 הבעיה שאנו מתכוונים לפתור

פיתוח כלי מדויק ככל שניתן על מנת לתת פתרון יעיל וחדשני לנושא סיווג מין לפי כתבי יד.
הכלי ייתן מענה גם למשתמשים רגילים הרוצים לבצע ניסויים על כתבי יד שונים, גם למשתמשים מתחומים אחרים כמו רפואה, פסיכולוגיה ועוד, וכן לחוקרים מהתחום המעוניינים לבצע ניסויים מעמיקים יותר שלהם הכלי יעניק מגוון רחב יותר של הגדרות ואפשרויות לסיווג.

4. יישום

4.1 ארכיטקטורה כללית – המערכת תכלול מספר רכיבים:

- עיבוד מקדים לנתונים.
- שימוש במודלים לביצוע סיווג.
- הערכת תוצאות הסיווג.
- ממשק משתמש.



4.2 דרישות

דרישות לפיתוח:

דרישה 1	
שם הדרישה	מיון מאגר נתונים לפי מין
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	יצירת סקריפט לעיבוד מקדים של תמונות כתבי היד ממאגר הנתונים ומיונם לפי זכר או נקבה.

דרישה 2	
שם הדרישה	עיבוד מקדים למאגר הנתונים
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	עיבוד מקדים לתמונות כתבי היד הכולל זיהוי אזור כתב היד ויצירת תמונה חדשה עם אזור הטקסט הרלוונטי, הסרת השורות הצהובות המיועדות לסימון אזור הכתיבה, והפיכת התמונה לגוון שחור לבן.

דרישה 3	
שם הדרישה	חלוקת מאגר הנתונים
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	חלוקת מאגר הנתונים לשלוש קטגוריות(Train, Test, Validation) עבור אימון המודלים.

דרישה 4	
שם הדרישה	חלוקת התמונות ל- Patches
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	יצירת סקריפט המבצע חלוקה ל- Patches של תמונות כתבי יד על ידי שימוש ב"חלון הזזה".

דרישה 5	
שם הדרישה	אימון מודלים
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	אימון מודלים קיימים בשיטת fine tuning על מאגר הנתונים.

דרישה 6	
שם הדרישה	הערכת המודלים
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	הערכת המודלים שאומנו באמצעות שימוש במאגר ה-Test, לפי מדד accuracy.

דרישה 7	
שם הדרישה	הערכת מודלים לפי שיטת ensemble
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	בחירת שלושה מודלים המובילים לפי מדד accuracy וביצוע הערכה לפי שיטת הכרעת הרוב.

דרישות פונקציונאליות:

דרישה 8	
שם הדרישה	העלאת תמונת כתב יד לסיווג
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	בעת לחיצת כפתור משתמש במערכת יכול לטעון תמונה המכילה סריקה של כתב יד על מנת לבצע סיווג מגדרי.

דרישה 9	
שם הדרישה	הצגת תמונת כתב היד
הרשאת משתמש	מערכת
עדיפות	בנונית
תיאור מפורט לאופן היישום	בעת סיום הטעינה, התמונה תוצג למשתמש מטעמי נוחות וכדי לוודא שהתמונה אכן נטענה בהצלחה.

דרישה 10	
שם הדרישה	בחירת שיטת הסיווג הרצויה
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	למשתמש תוצג תיבת Checkbox שבה יכול לבחור את שיטת העבודה, בין אם ירצה לבצע סיווג באמצעות מודל יחיד או באמצעות שילוב של שלושה מודלים.

דרישה 11	
שם הדרישה	בחירת סוג מודל/ים
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	בהתאם לבחירת שיטת הסיווג הרצויה, המשתמש יוכל לבחור מודל או מודלים לביצוע הסיווג.

דרישה 12	
שם הדרישה	ביצוע סיווג
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	למשתמש יוצג כפתור על מנת להתחיל את תהליך הסיווג בהתאם להגדרות שבחר.

דרישה 13	
שם הדרישה	הצגת סיכום נתוני הסיווג
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	בעת סיום פעולת הסיווג, למשתמש יוצג חלון עם סיכום הנתונים באופן כללי שיכללו: כמות patces שחולצו מהתמונה, כמות patces שסווגו לפי כל מגדר, תוצאת סיווג סופית לפי מודל יחיד או לחלופין תוצאת סיווג סופית לפי "שיטת הרוב" בין שלושה מודלים

דרישה 14	
שם הדרישה	הצגת נתוני סיווג באופן מפורט
הרשאת משתמש	מערכת
עדיפות	גבוהה
תיאור מפורט לאופן היישום	בעת סיום פעולת הסיווג, למשתמש יוצג סיכום מפורט של נתוני הסיווג הכוללים פירוט נתונים לפי כל patch שחולץ מהתמונה, הכולל: סיווג patch ובאיזו הסתברות בהתאם לסיווג של מודל יחיד או שלושה מודלים.

דרישה 15	
שם הדרישה	יצירת סרגל כלים לניווט
הרשאת משתמש	מערכת
עדיפות	בנונית
תיאור מפורט לאופן היישום	יצירת סרגל כלים לניווט נוח בין העמודים שהאתר יציע

דרישה 16	
שם הדרישה	עמוד הסבר על אופן ביצוע תהליך הסיווג
הרשאת משתמש	מערכת
עדיפות	בנונית
תיאור מפורט לאופן היישום	בעמוד זה יוצג פירוט מלא למשתמש לגבי העבודה המתרחשת מאחורי הקלעים החל מתהליך העיבוד מקדים ועד לביצוע הסיווג.

דרישה 17	
שם הדרישה	עמוד הסבר כללי על המודלים
הרשאת משתמש	מערכת
עדיפות	בנונית
תיאור מפורט לאופן היישום	בעמוד זה יוצג פירוט על מודל שהמשתמש יבחר להציג מידע לגביו, המידע יכלול תיאור כללי לכל מודל והארכיטקטורה.

דרישה 18	
שם הדרישה	About
הרשאת משתמש	מערכת
עדיפות	נמוכה
תיאור מפורט לאופן היישום	כפתור להצגת מידע כללי ויצירת קשר עם מפחתי המערכת.

4.3 אופי ומצב כללי של היישום

4.3.1 מצב קיים – המצב כיום בתחום המחקר בנושא סיווג מין לפי כתבי יד מתחדש מעת לעת, בוצעו מספר ניסויים בכל מיני כלים ושיטות שונות החל מבניית מודלים ושיפורם, שילוב של כלים ושיטות פסיכולוגיות לאיפון תכונות בכתב ושיטות חדשות כמו data distillation על מנת להרחיב מאגרי מידע קיימים.
התוצאות במחקרים וניסויים אלה נעים בין 60-80 אחוזי דיוק.

4.3.2 אופי המערכת המוצעת – המערכת תהיה מבוססת "למידה מעמיקה", איסוף ועיבוד נתונים הכוללים תמונות כתבי יד, אימון מודלים קיימים ומוכרים לעיבוד תמונה על הנתונים שנאספו, הערכת התוצאות, בחירת שלושת המודלים בעלי התוצאות הגבוהות ביותר ושילוב התוצאות שלהם על פי שיטת "הרוב".

4.3.3 אילוצים – לא קיימים אילוצים על המערכת.

4.4 תיחום חיצוני: משתמשים ומערכות משיקות

4.4.1 משתמשים – חוקרים ומשתמשים בתחומים כמו רפואה, פסיכולוגיה, סוציולוגיה, ותחומים רבים נוספים שמבוססים על זיהוי מגדרי.

4.5 תיחום פנימי: תת-מערכות ופונקציות ראשיות

- סקריפט לעיבוד מקדים עבור תמונות כתבי יד.
- סקריפט ליצירת מקטעים של חלקי טקסט(patches) להגדלת מאגר הנתונים.
- סקריפט לאימון מודל המקבל כקלט תיקיות (train, validation, test).
- סקריפט להערכת התוצאות.
- סקריפט לשילוב תוצאות שלושת המודלים בעלי אחוזי הדיוק הגבוהים ביותר.

4.6 מאגרי נתונים

מאגר של תמונות בפורמט JPEG המכילים טופס למילוי ידני של פרטים מקדימים כמו תאריך, מגורים, גיל ומין, וכן קטע טקסט קצר עם שורות בחלק התחתון של הדף על מנת להעתיק אותו לשם.
המאגר מכיל כ900 תמונות שיאחסנו בGoogle drive.

4.7 אבטחת מידע

המערכת לא דורשת אבטחת מידע.

4.8 דרישות מיוחדות

שרת חיצוני לאחסון נתונים ואפשרות לשימוש בGPU להרצת המודלים ובאחסונם.

5. טכנולוגיה ותשתית

5.1 ארכיטקטורת מערכת כללית – המערכת תעוצב בצורת אתר אינטרנט שיתאים להרצה בכל מחשב בעל גישה לאינטרנט.

5.2 חומרה מרכזית לפיתוח הפרויקט

- HP OMEN 2019
  - מעבד i7 9750h
  - זיכרון DDR4 16GB
- Lenovo legion 2018
  - מעבד i5 8300h
  - זיכרון DDR4 8GB

5.3 אחסון נתונים – הנתונים מאוחסנים לוקאלית על הדיסק הקשיח בשני המחשבים שנועדו לפיתוח ומגובים באמצעות אחסון נוסף בGoogle drive.

5.4 מערכת הפעלה – Windows 10.

5.5 כלי פיתוח

- JetBrains PyCharm – סביבת פיתוח משולבת לפיתוח תוכנות בעיקר בשפת פייתון, הסביבה מספקת שירותים כגון ניתוח קוד, דיבוג הקוד ותומכת בתכנות בסביבת אינטרנט באמצעות פלטפורמת Django, וכן ב- Data science.
- Google colab – כלי חינמי שפותח על ידי גוגל המספק שירותי ענן לשימוש בCPU וGPU חיצוניים למגוון רחב של פעולות ושימושים.

5.6 חומרה צד לקוח – מחשב סטנדרטי עם מערכת הפעלה וחיבור לאינטרנט.

5. מימוש

6.1 גורמים מעורבים

מנחות הפרויקט :

- ד"ר אירינה רבייב
- ד"ר מרינה ליטבק

מפתחים:

- אורי טביבי
- שון אסולין

6.2 תוכנית עבודה

6.2.1 שיטת עבודה

הפרויקט יפותח בשיטה האג'ילית, פיתוח הפרויקט יחולק למספר ספרינטים קצרים, כך שבתחילת כל ספרינט יינתנו משימות לביצוע על ידי מנחות הפרויקט, ובכל סוף ספרינט יתבצע דיווח והערכת התוצרים שהתקבלו.

מודל ניהול פרויקט אג'ילי, או Agile Project Management נחשב כחדשני ביותר, ומיועד לפיתוח מהיר של פתרונות. פירוש המילה הלועזית אג'יל (Agile) הוא "גמיש ומהיר."

מטרת המודל ליצור פתרון בזמן המהיר ביותר, מתוך הנחה כי קיים יותר ממחזור אחד של פיתוח, ובמהלך הפיתוח האפיון המוקדם עשוי להשתנות. ניהול פרויקט אג'ילי דורש מיומנות ויכולת גמישות, הן של הצוות המנהל את הפרויקט והן של הצוות המבצע, תוך חלוקת הפרויקט לחלקים קטנים מראש, ניסיון לפתחם, קבלת פידבק ופיתוח נוסף לאותו חלק בפרויקט.

ניסויים ותוצאות

לאחר שלב העיבוד המקדים ואימון המודלים, התבצעו ניסויים על קבוצת התמונות שהוגדרה כקבוצת הTest.

תחילה, ביצענו ניסויים על כתבי היד ללא חלוקה ל-patches, כלומר הכנסנו את תמונת כתב היד השלמה למודל לקבלת הסיווג.

לכל מודל ביצענו מספר ניסיונות שונים תוך כדי שינוי פרמטרים מסויימים כגון: מספר Epochs, שינוי ה-Learning rate ועוד, כדי לקבל תוצאות גבוהות ככל הניתן.

להלן התוצאות הסופיות שהתקבלו:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy
Vgg16	25	77.5%	70%	75%
Vgg19	20	75%	79%	75%
Xception	15	70%	77.7%	77%
EfficientNet	15	67%	72%	75%
NasNet	15	66%	60%	73%
Inception_V3	15	73%	60%	68%
DenseNet121	15	68%	65%	66%
DenseNet169	15	97%	78%	70%
Baseline	100	67%	65%	70%

כדי לשפר את אחוזי הדיוק, ביצעו ניסויים על כתבי היד עם חלוקה ל- patches.

לכל תמונה מקבוצת התמונות שהוגדרה כמאגר הTest מתבצעת חלוקה ל- Patches רבים ככל האפשר לפי אלגוריתם "חלון ההזזה", כל מודל בתורו מסווג כל patch באופן נפרד למחלקה שזיהה, זכר או נקבה, לפי תוצאות אילו מתבצע סיווג סופי למסמך הטקסט לפי שיטת הרוב, שבה המחלקה שקיבלה מספר הצבעות גבוה יותר, היא המחלקה שתקבע כסיווג הסופי.

לכל מודל ביצענו מספר ניסיונות שונים תוך כדי שינוי פרמטרים מסויימים כגון: מספר Epochs, שינוי ה-Learning rate ועוד, כדי לקבל תוצאות גבוהות ככל הניתן.

להלן התוצאות הסופיות שהתקבלו:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy	Test accuracy for text
Vgg16	12	77%	74%	79%	76%
Vgg16	19	78.5%	74%	80%	79%
Vgg19	18	77%	75%	72%	73.6%
Xception	15	73%	72%	79%	84%
Xception	20	75%	72%	79%	80%
Xception	40	78%	73%	79%	85%
EfficientNet	15	72%	74%	81%	84%
Inception_V3	15	75%	75%	75%	77%
DenseNet121	20	75%	73%	78%	74%
DenseNet169	20	76%	75%	78%	74%
NasNet	20	73.5%	73.5%	77%	84%
Baseline	100	74.81%	69.3%	80.02%	81.42%

- Test accuracy – אחוז הדיוק שהתקבל על ה-patches שנמצאים בתוך קבוצת ה-Test.
- Test accuracy for text – אחוז הדיוק שהתקבל עבור הסיווג הסופי של כל תמונת כתב יד.

ניסוי נוסף שנעשה על מנת לנסות לשפר את התוצאות הוא שימוש בסוגים שונים של אוגמנטציה ואלו הן:

- ביצוע סיבוב לתמונה בזווית בין 30 ל- 30- מעלות.
- ביצוע scaling עם פקטור רנדומלי בין 0.8 ל-1.2.
- הוספת רעש לתמונה.

תחילה ניסינו את השיטות האלה בנפרד ולאחר מכן ביחד, אך הניסוי לא הפיק תועלת משום שלא הצלחנו להשיג שיפור בתוצאות, יתרה מכך בחלק המקרים הייתה אפילו ירידה באחוזי הדיוק, ההסבר המשוער לכך הוא שהמאגר נתונים הוא עקבי, כלומר כל כתבי היד נסרקו תחת אותם תנאים ולכן גם קבוצת ה-test תהיה זהה לקבוצת ל-train, ולכן אם נוסיף שינוים באימון זה רק ישפיע לרעה על הזיהוי ב-test.

במקביל, התבצעו ניסויים על מאגר נתונים נוסף בשם ICDAR data set.

בשנת 2013 [12] התקיימה תחרות בהנחיית ICDAR, מטרת התחרות הייתה לאסוף כמה שיותר צוותים מהעולם על מנת לפתור את בעיית זיהוי מין הכותב על ידי תמונת כתב ידו בשיטות שונות שיפותחו על ידי הצוותים.

ICDAR סיפקו מאגר נתונים זהה לכל הצוותים הכולל כתבי יד של 475 כותבים, כל כותב סיפק ארבעה כתבי יד שונים, שניים בשפה הערבית ושניים בשפה האנגלית, כלומר 1900 כתבי יד סך הכל.

במאגר ישנם 221 כותבים ממין זכר כלומר סך הכל 884 טפסים משתי השפות, כאשר 442 בשפה האנגלית ו-442 בשפה הערבית.

כמו כן, ישנם 254 כותבים ממין נקבה כלומר סך הכל 1016 טפסים משתי השפות, כאשר 508 בשפה האנגלית ו-508 בשפה הערבית.

תחילה ביצענו ניסויים בהם המודלים אומנו על כתבי היד בשתי השפות ללא חלוקה ל-patches, להלן התוצאות:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy
Vgg16	20	67%	60%	52%
Vgg19	20	68%	60%	57%
Xception	20	67%	68%	58%
EfficientNet	15	62%	67%	60%
Inception_ResNet_V2	20	67%	55%	51%
Inception_V3	15	66%	63%	54%
DenseNet121	20	62%	66%	65%
DenseNet169	20	63%	66%	61%
ResNet50	20	73%	59%	52%
NasNet	15	63%	58%	60%
Baseline	100	55.63%	65.34%	55.68%

לאחר מכן ביצענו ניסויים בהם המודלים אומנו על כתבי היד בשתי השפות עם חלוקה ל-patches, להלן התוצאות:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy	Test accuracy for text	AVG log loss	Majority log loss	Log loss by id
Vgg16	19	67%	62%	62%	65%	0.62	0.71	0.6
Vgg19	19	67%	62%	62%	67%	0.63	0.74	0.62
Xception	20	70%	66%	50%	68%	0.55	0.57	0.54
EfficientNet	15	67%	68%	66%	67%	0.59	0.67	0.59
Inception_ResNet_V2	10	77%	67%	65%	75%	0.54	0.54	0.53
Inception_V3	15	70%	67%	66%	70.5%	0.56	0.57	0.56
DenseNet121	15	69%	67%	66%	69%	0.59	0.63	0.58
DenseNet169	15	70%	66%	66%	67%	0.56	0.6	0.56
ResNet50	20	75%	59%	59%	67%	0.61	0.64	0.6
NasNet	15	65%	65%	62%	66%	0.75	1	0.75
Baseline	100	64.64%	67.38%	64.92%	55.52%	0.7	1.4	0.78

לאחר מכן, ביצענו ניסויים בהם המודלים אומנו על כתבי היד עם חלוקה לפי שפות, להלן התוצאות עבור השפה האנגלית:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy	Test accuracy for text	AVG log loss	Majority log loss	Log loss by id
Vgg16	19	71%	57%	65%	70%	0.6	0.63	0.59
Vgg19	20	72%	59%	64%	69%	0.58	0.55	0.58
Xception	20	73.5%	68.5%	72%	75%	0.51	0.49	0.50
EfficientNet	15	68.5%	64.5%	68%	75%	0.58	0.64	0.58
Inception_ResNet_V2	10	82%	65%	65%	71%	0.59	0.63	0.58
Inception_V3	15	72%	65%	70%	72.5%	0.51	0.48	0.51
DenseNet121	15	71%	68%	70%	67.5%	0.56	0.59	0.55
DenseNet169	15	72%	65%	71%	71%	0.53	0.52	0.52
ResNet50	20	80%	57%	59%	67%	0.64	0.68	0.64
NasNet	15	68%	62%	67%	60%	0.67	1	0.67
Baseline	100	66.37%	66.06%	67.48%	52.46%	0.7	1.36	0.82

להלן התוצאות עבור השפה הערבית:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy	Test accuracy for text	AVG log loss	Majority log loss	Log loss by id
Vgg16	19	71%	65%	62%	60%	0.68	0.97	0.68
Vgg19	20	70%	63%	61%	66%	0.62	0.71	0.62
Xception	20	70%	70%	70%	63%	0.6	0.8	0.6
EfficientNet	15	68%	69%	67%	69%	0.58	0.64	0.59
Inception_ResNet_V2	10	81%	68%	65%	65%	0.6	0.64	0.6
Inception_V3	15	70%	71%	69%	68.5%	0.57	0.58	0.57
DenseNet121	15	69%	71%	70%	65%	0.62	0.8	0.63
DenseNet169	15	70%	71%	69%	64%	0.59	0.7	0.59
ResNet50	20	82%	59%	51%	51%	0.92	1.37	0.91
NasNet	15	66%	69%	65%	50%	0.71	1	0.72
Baseline	100	65.52%	68.5%	63.92%	61.45%	0.65	1.12	0.81

דירוג הקבוצות בתחרות התבצע על ידי חישוב Log loss, זהו מדד המראה כמה ההסתברות של הזיהוי קרובה לתשובה הנכונה ובהתאם לכך מצביע על איכות התוצאות.

- AVG log loss – המדד מחושב לפי ממוצע של הסתברות שהמודל חישב לכל Patch בנפרד.
- Majority log loss – במדד זה ההסתברות של דגימה להיות שייכת למחלקה מסוימת היא לפי רוב ה-patches שזוהו כשייכים למחלקה זו מתוך כל הכמות

שחולצה, לדוגמא: אם חולצו עשרה patches, ושבעה מתוכם זוהו כשייכים ל'נקבה', ההסתברות של כתב היד להיות שייך למחלקה 'נקבה' היא 0.7 (7/10).

- Log loss by id – ההסתברות של כותב מסוים להיות שייך למחלקה מסוימת, כל כותב סיפק ארבעה סוגים של כתבי יד, ולכן ההסתברות של כותב מסוים להיות שייך למחלקה מסוימת תלויה בממוצע ההסתברויות של כתבי היד שסיפק.

התוצאות שקיבלנו בהשוואה לתוצאות המתחרים בתחרות, מציבות אותנו בין עשרת הראשונים.

בנוסף בשנת [11] 2015 התקיימה תחרות נוספת בהנחיית ICDAR, התחרות התנהלה בצורה דומה לתחרות משנת 2013, אך בנוסף לבדיקות שהתבצעו באימון מודלים על שפה מסוימת וזיהוי על אותה שפה, נבדקה האפשרות המעניינת שבה אימון מודלים מתבצע על שפה מסוימת אך הזיהוי מתבצע על השפה האחרת, כלומר אם ביצענו אימון על השפה האנגלית, הזיהוי יתבצע על השפה הערבית ולהפך.

ICDAR סיפקו מאגר נתונים שונה מהתחרות של 2013, מאגר זה הוא תת מאגר של QUWI.

למשימת הזיהוי סופקו 300 דגימות כתב יד כקבוצת אימון, 100 עבור ולידציה ו100 עבור הטסט.

תחילה, אימנו את המודלים בשפה **האנגלית** ובחנו את תוצאות הזיהוי בשפה **האנגלית ובשפה הערבית**, להלן התוצאות:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy	Test accuracy for text – English	Test accuracy for text – Arabic
Vgg16	19	72.62%	65.32%	60.54%	66.5%	51%
Vgg19	20	72.35%	65.27%	60.24%	60%	52%
Xception	40	74.68%	69.90%	66.27%	74.5%	54.5%
EfficientNet	15	68.43%	66.73%	64.39%	68.5%	55%
Inception_ResNet_V2	10	84.92%	64.97%	62.72%	68%	54%
Inception_V3	15	72.74%	71.25%	64.96%	74%	54%
DenseNet121	15	71.88%	69.10%	64.22%	70.5%	50.5%
DenseNet169	15	71.46%	72.25%	65.83%	71.5%	50%
NasNet	15	72.19%	70.78%	66.05%	74%	55%
ResNet50	13	77.21%	66.83%	56.93%	65%	55%
Baseline	100	65.93%	67.90%	63.48%	61.5%	52%

לאחר מכן אימנו את המודלים בשפה **הערבית** ובחנו את תוצאות הזיהוי **בשפה האנגלית ובשפה הערבית**, להלן התוצאות:

Model name	Epochs	Train accuracy	Validation accuracy	Test accuracy	Test accuracy for text – English	Test accuracy for text – Arabic
Vgg16	19	72.47%	53.92%	60.62%	63.5%	56%
Vgg19	20	72.71%	63.05%	59.78%	60%	54.5%
Xception	40	72.20%	64.97%	63.20%	65.5%	63.5%
EfficientNet	15	69.66%	65.24%	66.49%	58.8%	66.5%
Inception_ResNet_V2	10	83.25%	72.58%	61.57%	64%	65%
Inception_V3	15	71.89%	63.89%	62.67%	69.5%	60.5%
DenseNet121	15	69.91%	62.58%	64.03%	64%	60.5%
DenseNet169	15	70.89%	65.32%	66.10%	61.5%	62.5%
NasNet	15	68.68%	60.61%	63.87%	61.5%	66%
ResNet50	13	78.67%	60.24%	58.79%	57%	57.5%
Baseline	100	67.29%	69.61%	63.24%	64.5%	60.5%

להלן הקטגוריות שהוגדרו בתחרות והתוצאה הגבוהה ביותר שהושגה:

2A – אימון מודלים וזיהוי מין על קבוצת Test בשפה הערבית; 65%

2B – אימון מודלים וזיהוי מין על קבוצת Test בשפה אנגלית; 60%

2C – אימון מודלים בשפה הערבית וזיהוי מין על קבוצת Test בשפה האנגלית; 63%

2D – אימון מודלים בשפה האנגלית וזיהוי מין על קבוצת Test בשפה הערבית; 58%

להלן התוצאות הטובות ביותר שהשגנו:

2A – 66.5%

2B – 74.5%

2C – 65.5%

2D – 55%

ניתן לראות כי הצלחנו לקבל תוצאות טובות יותר ב-3 מתוך 4 הקטגוריות.

ניסוי לבדיקת טעויות אנוש:

בשלב האחרון, לאחר ביצוע וסיכום התוצאות באמצעות כלים מעולם ה- Deep learning, ביצענו מחקר על ההשוואה בין טעויות המודלים לטעויות אנוש, לשם כך ריכזנו ארבעה שאלונים מקוונים.

השאלונים כללו 70 תמונות כתבי יד בשפה בעברית שמהוות את קבוצת ה- test שנבדקה גם על ידי המודלים, הקבוצה בסך הכל חולקה ל18, 18, 17, 17 עבור שאלונים 1-4 בהתאמה.

כל משתתף התבקש לחזות את מין הכותב של דוגמאות הטקסט שהוצגו לפניו בשאלון, בסיום השאלון המשתתף יכל לראות את התשובות שלו ביחס לתשובות הנכונות.

להלן סיכום התוצאות:

שאלון מספר 1 – 166 משתתפים, אחוז דיוק ממוצע – 62.3%.

שאלון מספר 2 – 109 משתתפים, אחוז דיוק ממוצע – 63.2%.

שאלון מספר 3 – 89 משתתפים, אחוז דיוק ממוצע – 73.9%.

שאלון מספר 4 – 86 משתתפים, אחוז דיוק ממוצע – 70.7%.

סך הכל ממוצע כללי 67.4%.

ניתן לראות באופן ברור כי תוצאות המודלים היו טובות בהרבה מתוצאות בני האדם.

משקלים רנדומליים (Random weights):

כחלק מניסיון שיפור תוצאות המודלים, בדקנו כיוון נוסף על ידי שינוי פונקצית המשקל בעת טעינת המודל.

המודלים שאיתם השתמשנו במהלך המחקר הם מודלים pre-trained על מאגר תמונות בשם Image-net, כך שהמשקלים שנטענו במודל היו על Image-net.

ניסוי זה כלל שינוי פונקצית המשקל מ- Image-net ל-random weights, על ידי שינוי השדה בפונקציה ל- None, כך המודל נטען לאימון ללא משקלים.

תהליך העבודה היה זהה, הרצנו את המודלים שוב על מאגר הנתונים HHD_gender בשפה העברית, התוצאות היו נמוכות יותר מהתוצאות שקיבלנו בתחילת הפרויקט.

נספחים

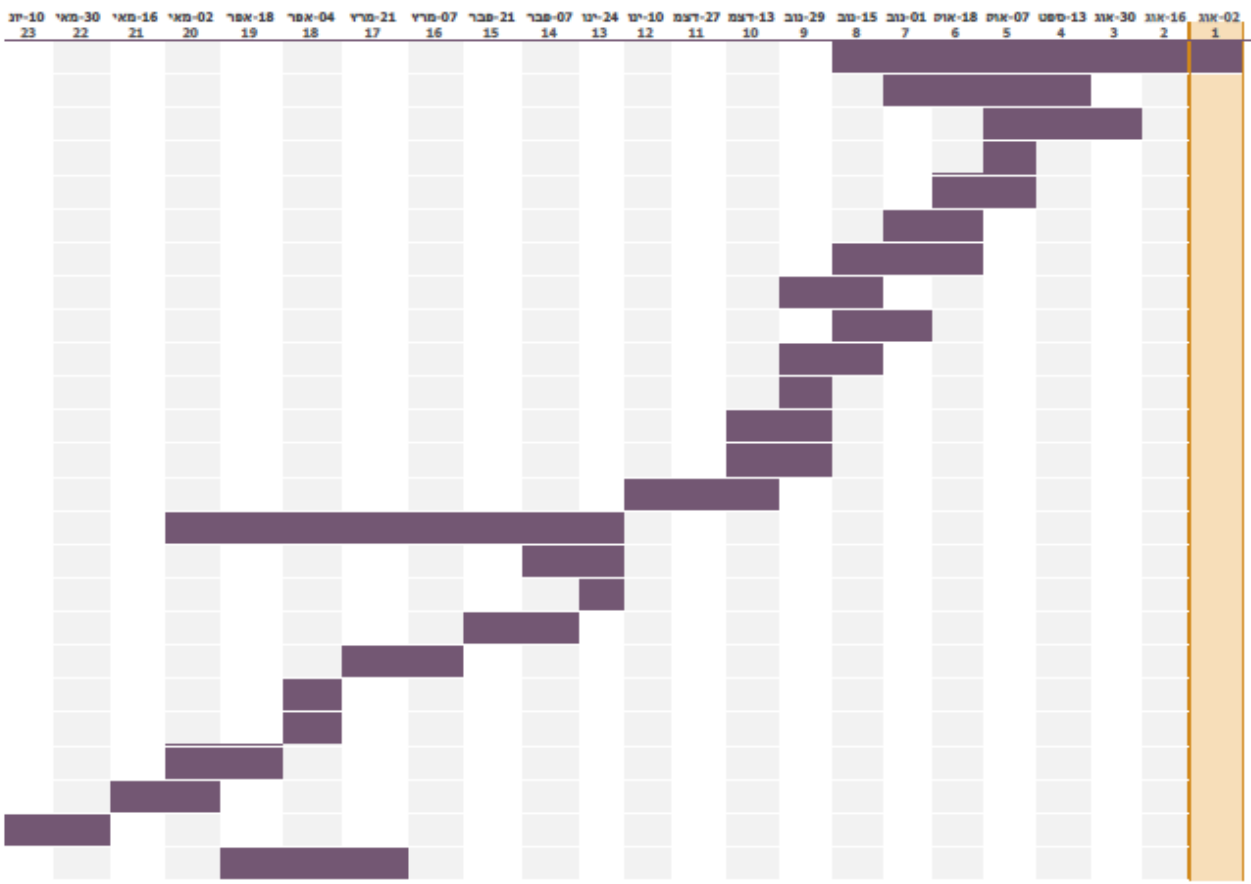
תוכנית עבודה

כדור הקופה כדי לסמן אותה בצד שמאל. הסקרא מתאר את סוגי התרשימים.

סימון תקופה: 1

משך התוכנית | התחלה בפועל | % ביצוע | בפועל (מעבר לתוכנית) | % ביצוע (מעבר לתוכנית)

פעילות	התחלת התוכנית	משך התוכנית	התחלה בפועל	משך בפועל	אחוז ביצוע
למידת וחקר הנושא	1	8	1	8	100%
כתיבת סקר ספרות	4	4	4	4	100%
Preprocessing	3	3	3	3	100%
בניית מודל לאימון	5	1	5	1	100%
ביצוע ניסויים	5	2	5	2	100%
ביצוע סגמנטציה לתמונות	6	2	6	2	100%
אימון מודלים לעיבוד תמונות	6	3	6	3	100%
חערכת תוצאות	8	2	8	2	100%
אימון מודלים על ICDAR	7	2	7	2	100%
חערכת תוצאות	8	2	8	2	100%
ביצוע שילוב בין מודלים	9	1	9	1	100%
כתיבת מסמך איפיון וייזום	9	2	9	2	100%
כתיבת מסמך דרישות	9	2	9	2	100%
למידת Flask	10	3	10	3	100%
פיתוח מערכת משתמש לחיזוי תמונה	13	8	13	8	100%
פיתוח תפריט הגדרות למשתמש	13	2	13	2	100%
פיתוח העלאת תמונה למערכת	13	1	13	1	100%
העלאת מודלים והתממשקותם למערכת	14	2	14	2	100%
פיתוח עמוד לחצגת תוצאות חיזוי התמונה	16	2	16	2	100%
פיתוח עמוד הסבר כללי למבנה המערכת	18	1	18	1	100%
פיתוח עמוד הסברים על המודלים	18	1	18	1	100%
בדיקות מערכת	19	2	19	2	100%
עריכת פוסטר פרוייקט לועדה סופית	20	2	20	2	100%
עריכת מצגת פרוייקט מסכמת	22	2	22	2	100%
סיום ספר פרוייקט	17	3	17	3	100%



מילון מונחים

מונח	הגדרה
למידה מעמיקה	הלמידה העמוקה (Deep Learning) היא תחום מחקר בעולם המחשבים וספציפית בתחום "למידת המכונה" שמניח שהמחשב יכול ללמוד וללמד את עצמו, ממש כמו המוח האנושי. מטרתו הברורה של התחום הזה היא ליצור חיקוי ממוחשב של פעולת המוח האנושי.
רשת נוירונים	רשתות נוירונים (Neural Networks) הן רשתות מחשבים מתקדמות שמחקות את החשיבה האנושית.
patches	חלקים קטנים המחולצים מתוך תמונת טקסט שלמה במטרה להרחיב את מאגר נתונים
עיבוד מקדים	שלבים בעיבוד הנתונים לפני אימון המודל, השלבים יכולים לכלול ניקוי, איסוף, עיבוד ומידול המידע
הערכת תוצאות	השלב שלאחר אימון המודלים, קבלת התוצאות והסקנת המסקנות
זיקוק נתונים data distillation	הרעיון הוא לאמן מודל באמצעות כמות גדולה מאוד של נתונים, לבצע סיווג ואז לאמן שוב את המודל באמצעות התגיות החדשות שנוצרו, עם זאת ישנה בעייתיות של מידע חסר חשיבות שנוצר מאימון מודל על התחזיות שהוא עצמו מייצר.
ensemble	ביצוע שילוב בין מודלים על מנת להעלאות את אחוז הדיוק, תוצאת הסיווג תתבצע לפי שיטת "הרוב" בין תוצאות סיווג המודלים.
אוגמנטציה	טכניקה להכפלת כמות המידע על ידי יצירת מאגר חדש על בסיס המאגר הקיים עם שינויים בהתאם למספר שיטות.

ניהול סיכונים

סיכון	רמת פגיעה	הסתברות	ציון משוקלל	דרך פתרון
מודלים עם אחוזי דיוק נמוכים מהמצופה	5	3	15	ביצוע אימונים נוספים וניסויים רבים על מנת לשפר ולייעל ככל הניתן את אחוזי הדיוק של המודלים
העלאת תמונה לסיווג באיכות שתפגע בביצוע הסיווג	5	3	15	מתן הסבר מפורט טרם העלאת התמונה בצורה שתאפשר למשתמש להבין באיזה פורמט ובאיזו איכות עליו להעלות את התמונה כך שלא תפגע בביצוע בסיווג.
חוסר כוח חישוב בשרת החיצוני	4	2	8	ביצוע ניסויים מקיפים ובדיקת מקרי קצה על מנת להבטיח כי לא יהיה מצב של חוסר כוח עיבוד לשרת החיצוני.
אי עמידה בזמנים	5	4	20	עמידה בצמוד לתוכנית העבודה שנכתבה. ביצוע מפגשים דו-שבועיים עם מנחות הפרויקט במטרה לפקח על ביצוע המשימות שניתנו.
חוסר ידע בתכנות ובממשק.	5	3	15	התחלת עבודה מוקדמת כדי שיהיה זמן לשאלות וללמידה. התייעצות עם אנשי מקצוע, מנחות הפרויקט ועם אנשים שכבר עבדו בתחום.
חוסר ניסיון בנושאים הקשורים לפיתוח ויזמות.	3	5	15	עבודה באופן שותף מול מנחות הפרויקט.
פיתוח ממשק לא נוח למשתמש	4	2	8	פיתוח ממשק בתוספת כפתורים נגישים והסברים נרחבים על אופן ביצוע כל פעולה
עבודת צוות כושלת	3	2	6	כל חלק מכתיבת העבודה יעשה בשיתוף פעולה בין חברי הקבוצה. כל דבר שייכתב יועבר לשאר וכך תמיד יהיה מעקב אחר ההתקדמות וכולם יהיו שותפים.
שינוי לא מתוכנן בתוכנית העבודה שלא במסגרת הלו"ז שנקבע	3	3	9	שינויים שידרשו בביצוע המשימות בצורה לא מתוכננת מראש יפגעו בעמידה בלו"ז לפיתוח הפרויקט ולכן על כל שינוי שיידרש הצוות יבצע הערכה מחדש ותיקון לתוכנית העבודה בצורה שלא תפגע בסיום הפרויקט.

ביבליוגרפיה

Public handwriting datasets:

QUWI - https://handwriting.datasetsonline.com/

ICDAR2013 –

https://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting/data

KHATT - http://khatt.ideas2serve.net/KHATTAgreement.php

IAM - http://www.fki.inf.unibe.ch/databases/iam-handwriting-database/download-the-iam-handwriting-database

References:

1. Al Maadeed, S., Hassaine, A.: Automatic prediction of age, gender, and nationality in offline handwriting. EURASIP J. Image Video Process. **2014**(1), 10 (2014).
2. Liwicki, M., Schlapbach, A., Loretan, P., Bunke, H.: Automatic detection of gender and handedness from on-line handwriting. In: Proceedings of the 13th Conference of the Graphonomics Society, pp. 179–183 (2007).
3. Siddiqi, I., Djeddi, C., Raza, A., Souici-Meslati, L.: Automatic analysis of handwriting for gender classification. Pattern Anal. Appl. 18(4), 887–899 (2015).
4. Radosavovic I, Doll´ar P, Girshick R, Gkioxari G, He K.: Data Distillation: Towards Omni-Supervised Learning. Facebook AI Research (FAIR). CVPR (2018).
5. Upadhyay S, Singh J.: Determination of Sex Through Handwriting Characteristics. IJCRR Section: General Science Sci. Journal Impact Factor 4.016 ICV: 71.54 (2017).
6. Topaloglu M, Ekmekci S.: Gender detection and identifying one's handwriting with handwriting analysis. Expert Systems With Applications 79 (2017) 236–243.
7. Xie Q, Luong M.T, Hovy E, V. Le Q.; Self-training with Noisy Student improves ImageNet classification. CVPR 2020 open access.
8. Gattal A, Djeddi C, Bensefia A, Ennaji A.: Handwriting Based Gender Classification Using COLD and Hinge Features. ICISP 2020: Image and Signal Processing pp 233-242.
9. Morera Á, Sánchez Á, Vélez J.F, Moreno A.B.: Gender and Handedness Prediction from Offline Handwriting Using Convolutional Neural Networks. Hindawi Complexity Volume 2018, Article ID 3891624, 14 pages.
10. Illouz E, David E, Netanyahu N.: Handwriting-Based Gender Classification Using End-to-End Deep Neural Networks. ICANN 2018: Artificial Neural Networks and Machine Learning – ICANN 2018 pp 613-621.

11. Chawki D, Somaya Al-Maadeed; Abdeljalil G; Imran S; Labiba Souici-Meslati.: ICDAR2015 competition on Multi-script Writer Identification and Gender Classification using 'QUWI' Database - Competitions @ 2015 13th International Conference on Document Analysis and Recognition (ICDAR).
12. Abdelâali H; Somaya Al Maadeed; Jihad A; Ali J.: ICDAR 2013 Competition on Gender Prediction from Handwriting - 2013 12th International Conference on Document Analysis and Recognition.

www.sce.ac.il

קמפוס באר שבע

ביאליק 56, באר שבע 84100

קמפוס אשדוד

ז'בוטינסקי 84, אשדוד 77245