# **סיווג חזה אוויר בריאות באמצעות רשתות עצביות עמוקות**

# **Pneumothorax Classification Using Deep Neural Networks**

מאת אילת ג'יבלי וטהר צביטמן

## **תקציר** במחקר זה ניסינו לאמן רשת נוירונים שתצליח לזהות מצב של חזה אוויר בתצלומי רנטגן של ריאות.

חזה אוויר הינו מצב רפואי היכול להתפתח למסכן חיים במידה ואינו מאובחן ומטופל בזמן. אבחון של חזה אויר נעשה רק על ידי קרדיולוג מומחה, עקב כך האבחון של מטופלים לעיתים נדחה זמן רב בשל מחסור בקרדיולוגים מומחים. רשת נוירונים שתצליח לזהות חזה אוויר בצילומי רנטגן באחוזי דיוק גבוהים יכולה להפוך לכלי עזר רפואי אשר יקל על העומס במערך הרפואי ויאפשר לתת קדימות בטיפול לחולים אשר הרשת זיהתה אצלם חזה אוויר.

השתמשנו בדאטה המכיל 2027 תמונות של צילומי חזה אשר צולמו במרפאה של קרדיולוגים וסווגו על ידם ולאחר עיבוד הדאטה קיבלנו 3317 תמונות בהן השתמשנו על מנת לאמן רשתות השונות. הצלחנו להגיע לאחוזי דיוק מרשימים של 91% בעזרת רשת המבוססת על resnet18. אלו אחוזי הדיוק הטובים ביותר שמצאנו עבור רשתות שאומנו על הדאטה שלנו, תוצאות טובות יותר התקבלו רק עבור מאגרי מידע אחרים וגדולים משמעותית.

**מבוא**  
חזה אוויר[1], הוא מצב רפואי בו מצטבר אוויר בין קרומי האדר המחברים את הריאות לסרעפת. במצב זה נמנע מהריאה להכניס ולהוציא אוויר בצורה הרגילה ואף ייתכן מצב

שבו נוצר לחץ על הלב כתוצאה מהאוויר המצטבר. במידה וחזה אוויר לא אובחן בזמן, הוא יכול להתפתח למצב מסכן חיים.

חזה אוויר יכול להיגרם ממספר סיבות: 1. כאשר ישנה פציעה הגורמת לקרע במשטח הריאה או דופן החזה ומאפשרת לאוויר להיכנס לחלל האדר. 2. כתוצאה ממחלת ריאה, נוצרת דליפת אוויר מהריאה עצמה לחלל האדר. 3. באופן ספונטני וללא סיבה ישירה.

חזה אוויר מאובחן בדרך כלל על ידי קרדיולוג מומחה בעזרת צילום חזה, ולעיתים יכול להיות מאוד קשה לאבחון. לרוב ישנו מומחה אחד שעליו לעבור על עשרות סריקות. שימוש באלגוריתם AI מדויק לזיהוי חזה אוויר בצילומי ריאות יהיה שימושי בהרבה תרחישים קליניים. ניתן להשתמש באלגוריתם כזה על מנת לתת עדיפות למקרים מסוימים, חוות דעת נוספת ועזרה לקרדיולוג, או לאפשר אבחון על ידי רופאים שאינם קרדיולוגים במקרה והדיוק מאוד גבוהה.

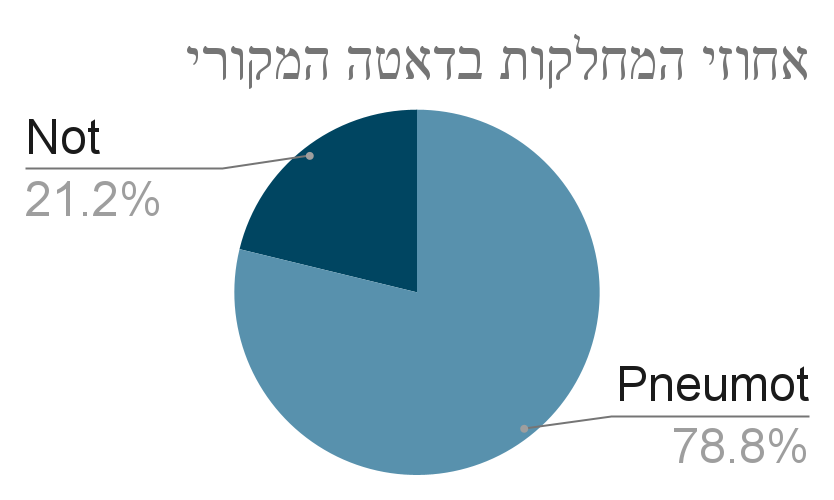
שימוש באלגוריתמי למידה עמוקה מהווה פריצת דרך בתחום. לפני כן, השתמשו באלגוריתמי למידת מכונה קלאסיים על מנת לאבחן את חזה האוויר. החסרון היחיד של שימוש בלמידה עמוקה הוא הצורך בהרבה דאטה מאובחן ומסודר שעליו מאמנים את המודל. בעיה זאת נפתרה בעזרת הזמינות של מאגרי מידע גדולים בפלטפורמות חינמיות שונות, כגון kaggle, Google-dataset-search. בעזרת פלטפורמות אלו מפתחים מכל העולם יכולים לתרום לאוטומציה של תחומים רבים. בפרויקט זה ניסינו לאמן רשת המסווגת תמונות CT של חולים ממאגר מידע[2] מאתר kaggle שנלקח מתוך תחרות[3] שפורסמה באתר ב2019 "SIIM-ACR Pneumothorax Segmentation". ביצענו ניסויים רבים וסיכמנו את ההשפעות של רכיבים שונים על התוצאות.

## **עבודות קודמות** רוב מאגרי המידע המכילים תמונות CT של החזה הם בלתי מאוזנים, אינם מיועדים ספציפית למשימה מסוימת ומכילים מעט מאוד מקרים של חזה אוויר. בניגוד לכך, מאגר המידע שבחרנו מיועד רק למטרה זו, ולכן קטן משמעותית. ישנם מספר שיטות לאיזון המידע, כאשר מרבית המאמרים נקטו בשיטה האלגוריתמית. בפרויקט זה אנו נקטנו בגישה שונה - לטפל בבעיה בעזרת שינוי המידע עצמו. על מנת לסווג את התמונות, למידה עמוקה היא השיטה הפופולרית שנותנת את התוצאות הטובות ביותר. בתוך למידת עמוקה, נפוץ השימוש במודלים ידועים וקיימים שכבר אומנו על מאגרי מידע גדולים (כגון VGG16, ResNet ו DenseNet) , בשל החסכון הרב בזמן האימון. גם בפרויקט זה, בשל הוא נעשה על המחשב הביתי, התמקדנו בשיטה זו ולא בפיתוח רשת חדשה עם המון שכבות. במרבית המחקרים וכן במחקר זה, לא היה התייחסות לפרמטרים נוספים על החולה כגון גילו, מינו וכד'.

מאמרים לציון:

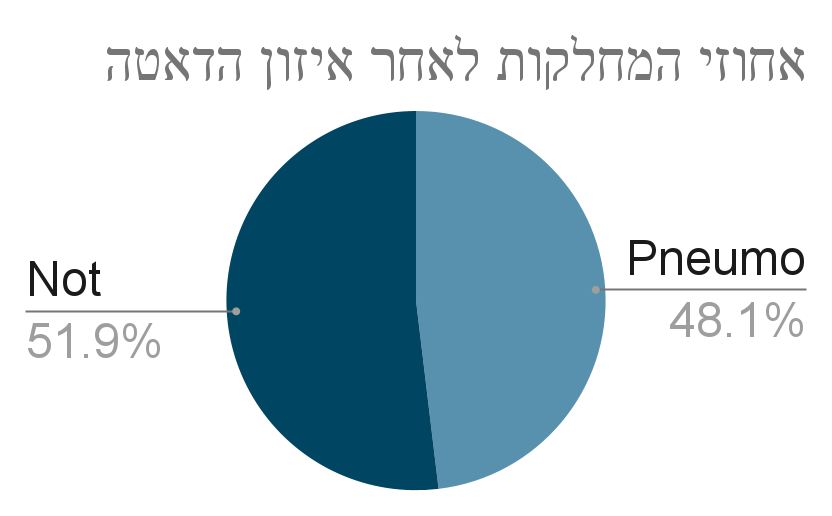
בשנת 2018 פורסם מאמר[4] הבוחן את שלושת רשתות ה resnet ברזולוציות שונות על מאגר המידע הגדול NIH. אחוז ההצלחה של המודל הטוב ביותר היה 91.1. בשנת 2019 פורסם מאמר[5] המשתמש ב YOLO Darknet19 על מאגר מידע פרטי ולא מאוזן (ללא הסבר על אופן איזון הדאטה). אחוז ההצלחה של המודל הטוב ביותר היה 90.5. בשנת 2020 פורסם מאמר[6] המציע פתרון בעזרת Densenet עם אחוז הצלחה 93.45 על מאגר מידע פרטי. בשנת 2021 פורסם מאמר[7] המציע פתרון בעזרת ResNet עם אחוז הצלחה 94.4. בנוסף, פורסם מאמר[8] הסוקר את כל הרשתות שפורסמו עד לאותה שנה לבעיה.

## **נתונים** מאגר המידע של הפרויקט[2] זמין באתר kaggle ומכיל 2027 תמונות רנטגן של צילומי חזה מסווגים לריאה בריאה או חולה ברזולוציה של 1024x1024. לפני הרצת הרשת, ביצענו עיבוד מקדים של הנתונים שהתחלק לכמה שלבים.

בשלב הראשון, חישבנו את האחוזים של כל מחלקה בדאטה, על מנת לבדוק האם הדאטה מאוזן.

ניתן לראות מהתוצאות כי הדאטה מוטה מאוד. האחוזים הגבוהים של המחלקה Pneumothorax יגרמו ל overfitting שכן חיזוי תמידי של 1 ייתן אחוזי דיוק גבוהים.

בשלב השני בדקנו כמה פעמים צריך לשכפל את המחלקה Not pneumothorax על מנת לאזן את הדאטה כמה שאפשר. לאחר שהרצנו מספר בדיקות שבכל אחת מהן ניסינו לשכפל את Not pneumothorax מספר שונה של פעמים, גילינו ששכפול של הדאטה במחלקה 4 פעמים יביא לתוצאות הטובות ביותר. לכן, בשלב הבא, שכפלנו כל תמונה בדאטה ויצרנו לה 3 העתקים זהים, בנוסף לעותק המקורי.

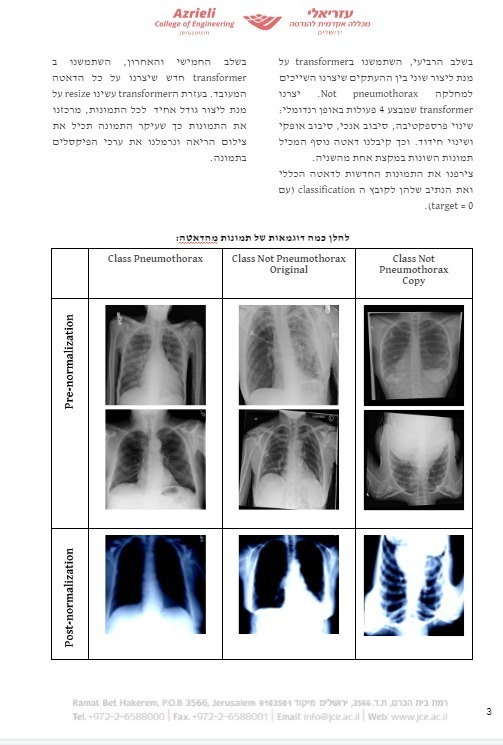


בשלב הרביעי, השתמשנו בtransformer על מנת ליצור שוני בין ההעתקים שיצרנו השייכים למחלקה Not pneumothorax. יצרנו transformer שמבצע 4 פעולות באופן רנדומלי: שינוי פרספקטיבה, סיבוב אנכי, סיבוב אופקי ושינוי חידוד. וכך קיבלנו דאטה נוסף המכיל תמונות השונות במקצת אחת מהשניה.

צירפנו את התמונות החדשות לדאטה הכללי ואת הנתיב שלהן לקובץ ה classification (עם target = 0). קישור לקוד המקור בgithub מצורף בנספחים [9].

בשלב החמישי והאחרון, השתמשנו ב transformer חדש שיצרנו על כל הדאטה המעובד. בעזרת הtransformer עשינו resize על מנת ליצור גודל אחיד לכל התמונות, מרכזנו את התמונות כך שעיקר התמונה תכיל את צילום הריאה ונרמלנו את ערכי הפיקסלים בתמונה.

להלן כמה דוגמאות של תמונות מהדאטה:

****

## **שיטות**

ננסה לחקור 2 שיטות ידועות ונפוצות בלמידה עמוקה. הראשונה, רשת CNN - רשת של שכבות קונבולוציה המיועדת ללמוד את התכונות של הדאטה. השניה, שיטת transfer learning. מאפייני הרשתות שהגדרנו:

הדאטה יחולק לשלושה קבוצות - train 80%, valid 10%, test 10%. כפונקצית הloss בחרנו את BCEWithLogitsLoss והתמקדנו בסיווג בינארי - יש חזה אוויר או לא. מכיוון שבתוך פונקצית loss זאת מופעל sigmoid לא הוספנו את sigmoid לסוף הרשת. ראינו במאמרים שונים [7] אשר חקרו דאטה זהה או דומה לשלנו, כי האופטימיזר Adam הביא באופן עקבי לתוצאות הטובות ביותר, לכן בחרנו להשתמש בו בכל הניסויים שביצענו. את בחינת דיוק הרשת ביצענו בעזרת מדד accuracy בינארי על קבוצת הtest (על ידי מס' התמונות שסווגו נכון חלקי סך התמונות בtest) והצגת הconfusion matrix. הקוד נכתב בפיית'ון[9] בעזרת ספרית pytorch.

בשלב הראשון הרצנו את הדאטה על רשת פשוטה המכילה 2 שכבות קןנבולוציה ו3 שכבות ליניאריות fully connected. השתמשנו בפונקציית relu כפונקציית אקטיבציה. הגענו לתוצאות סבירות אך לא מספקות עם accuracy = 72.53012 על הטסט. הסיבה לאחוזים הלא גבוהים היא שבמודל נוצר overfitting גבוה.

בשלב השני החלטנו לנסות את שיטת transfer learning עם רשתות מוכרות לסיווג של תמונות ולחקור את התוצאות עם פרמטרים שונים. בחרנו ברשתות הבאות:

vgg16/18 , denseNet161 , resnet50/18. לכל רשת כזאת הוספנו בסופה שכבות ליניאריות עם פונקציית relu ביניהן על מנת שהפלט יהיה ליניארי והרצנו את השכבות עם פרמטרים שונים. הפרמטרים שעליהם עשינו ניסויים על מנת לחקור את הרשתות הם:

batch size, epochs number, learning rate.

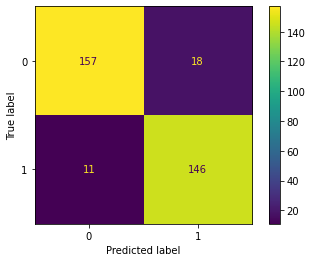
## **תוצאות של ניסויים**

לאחר מספר ניסויים הבנו שהגישה של רשת CNN לא תעבוד עבור הפרויקט שלנו. על מנת לשפר את הרשת צריך כוח חישובי גדול משמעותית מזה שלרשותינו כדי שנוכל לשמור את כל הפיצ'רים שהרשת צריכה. בנוסף, רשת הCNN יצרה את הOverfitting המשמעותי ביותר שהתקבל בכל הניסויים שהרצנו- 98% דיוק לtrain ו77% דיוק ל validation. לאחר מכן עברנו לגישה השנייה והשתמשנו ברשתות קיימות.

הרצנו את הרשתות שנבחרו עם פרמטרים שונים. במרבית מהרשתות האלו אכן קיבלנו overfitting נמוך בהרבה ממה שקיבלנו ברשת הCNN ולכן בחרנו להמשיך להתמקד בגישה זו תוך שינוי של הפרמטרים והוספת שכבות שונות לרשתות.

בשביל כיוונון השכבה האחרונה ברשתות הקיימות ניסינו 3 אפשרויות: אפשרות ראשונה- שכבה לינארית אחת. אפשרות שניה- שלוש שכבות לינאריות כשבינהם פונקציה relu כפונקצית אקיטבציה. אפשרות שלישית- שכבה לינארית, dropuot של 0.5 אחוז, שכבה לינארית, relu, שכבה לינארית.

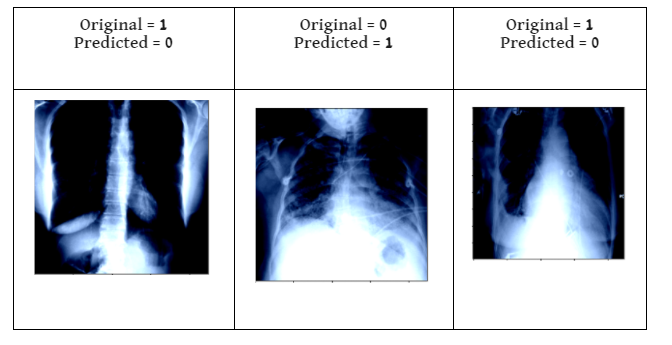
הרשת עם אחוזי הדיוק הגבוהים שהתקבלה ביותר היא resnet18 עם 3 שכבות ליניאריות וdropout. הרשת השיגה אחוזי דיוק של 91.26 על קבוצת הtests. להלן ה confusion matrix של הרשת:

ניתן לראות שה false negative הוא עם אחוז נמוך יותר מה false positive וזה אכן המצב הנדרש כאשר עובדים עם data רפואי.

## **טבלת סיכום של הניסויים הנבחרים:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 72.53012 | 1.5306 | 77.0385 | 0.0554 | 98.5 | Adam | 16 | 0.001 | 8 | **CNN basic model** |
| 84.45783 | 0.3356 | 84.5288 | 0.3633 | 83.7788 | Adam | 8 | 0.001 | 20 | **vgg16 + linear layer** |
| 90.66265 | 0.3265 | 86.7500 | 0.1752 | 92.5926 | Adam | 100 | 0.001 | 20 | **DenseNene 161 + linear layer** |
| 88.25301 | 0.3116 | 84.0000 | 0.2225 | 90.2222 | Adam | 300 | 0.01 | 17 | **DenseNene 161 + linear layer** |
| 88.25301 | 0.2918 | 86.2500 | 0.3188 | 85.5185 | Adam | 100 | 0.001 | 20 | **resnet 50 + linear layer** |
| 87.34939 | 0.3378 | 84.0000 | 0.3591 | 83.8182 | Adam | 250 | 0.01 | 10 | **resnet 50 + linear layer** |
| 91.26506 | 0.2971 | 86.0000 | 0.2710 | 88.4815 | Adam | 100 | 0.001 | 20 | **resnet 18**  **+ 3 linear layers + dropout** |
| 89.4578 | 0.3534 | 84.7500 | 0.2402 | 89.4815 | Adam | 100 | 0.001 | 20 | **resnet 18**  **+3 linear layers** |
| 88.8554 | 0.3919 | 88.3333 | 0.0073 | 100.00 | Adam | 150 | 0.0001 | 10 | **Vgg19 + 3 linear layers** |
| 89.1566 | 0.2934 | 86.5000 | 0.2856 | 87.2143 | Adam | 200 | 0.001 | 17 | **resnet 50 + 3 linear layers + dropout** |

## **ניתוח שגיאות**

נרצה לבחון שלושה מקרים רנדומליים שבהם הרשת הוציאה סיווג שגוי.

בשל כך שאיננו קרדיולוגיות או בעלות ידע כלשהו ברפואה, איננו יודעות לאבחן בעצמנו את התמונות ולהבין מה יכול לגרום לבעיה. השערה סבירה לגורם לכשלון של הרשת, היא כי בכל התמונות הללו יש "רעש" לבן המפריע לחיזוי. על מנת לשפר את הכלי בעתיד, אפשר להוסיף שלב של עיבוד תמונה מקדים על מנת לבודד רק את החלק המתאים בריאה ולבצע הכנה מתאימה של התמונות (למשל חידוד הedges). תהליך זה צריך להתבצע לאחר ייעוץ יסודי עם קרדיולוגים לאופן האבחון של המחלה על מנת להבין את המקומות בהם הרשת נופלת ובאיזה חלקים של התמונה להתמקד.

**מסקנות**

הרשתות השונות שאימנו הגיעו לאחוזי דיוק מרשימים. מאחר ומדובר במשימת סיווג קשה המצריכה רקע רפואי מתאים, רשת בעלת דיוק של 91% יכולה בהחלט להפוך לכלי רפואי שמיש (למשל על מנת לקבוע את דחיפות הטיפול הרפואי ולאפשר מתן קדימות למטופלים שהרשת קבעה כי יש להם חזה אויר). עם זאת, ניתן להגיע לאחוזי דיוק גבוהים יותר באמצעות אימון של הרשת על data גדול יותר.

לסיכום, על אף התוצאות הטובות אליהן הגענו, ניתן להשתמש במחקר זה כעבודה מקדימה ולהמשיך לפתח את הרשת באמצעות datasets שונים וכוח מחשוב גדול יותר.