

האוניברסיטה הפתוחה

המחלקה למתמטיקה ולמדעי המחשב

סקירה השוואתית של שיטות לסיווג מחלות ריאה מצילומי רנטגן

עבודה מסכמת זו הוגשה כחלק מהדרישות לקבלת תואר

מוסמך למדעים M.Sc. במדעי המחשב

באוניברסיטה הפתוחה

החטיבה למדעי המחשב

נכתב על-ידי

טל כפיר

העבודה הוכנה בהדרכתה של ד"ר מיה הרמן

מאי 2024

1. תקציר

מחלות ריאה רבות פוגעות באיכות החיים של בני האדם ויכולות לגרום לתסמינים חמורים ולעיתים למוות. מחלות ריאה ניתנות לאבחנה בדרכים שונות וביניהן צילום רנטגן של בית החזה. צילום רנטגן הוא כלי מרכזי, זמין, זול ונוח לאבחון מחלות ריאה כגון קורונה, דלקת ריאות נגיפית או ויראלית, עכירות ריאה, סרטן ריאה ומחלות ריאה נוספות. האבחון מתבצע על ידי רופאים מומחים המסוגלים לפענח את צילום הרנטגן של בית החזה, ולזהות בו דפוסים המעידים על מחלת הריאות.

תהליך האבחון והסיווג הינו ידני, ארוך ומורכב אשר דורש מומחיות רפואית ולעיתים אינו יכול להתבצע בזמן אמת וזאת בשל אי נוכחותו של רופא בעל מומחיות מתאימה. לפיכך, נדרשת מערכת אוטומטית לסיווג מחלות ריאה על סמך צילום רנטגן של בית החזה היכולה לסווג את התצלום לסוג המחלה הנכון. מערכת כזו תאפשר התאמת טיפול מתאים לנבדק ובהמשך מעקב אחר יעילות הטיפול ומיגור המחלה. במהלך השנים האחרונות, לאור התפתחות היכולות החישוביות וצמיחת תחום הלמידה העמוקה החלו להשתמש בלמידה עמוקה לצרכי זיהוי וסיווג צילומי רנטגן. בשנת 2020 פרצה מגפת הקורונה בקנה מידה עולמי. הקורונה היא מחלה מדבקת, הפוגעת בכלי הנשימה וגבתה את חייהם של מיליונים ברחבי העולם. התפרצות הקורונה תרם לצורך לזהות במהירות את המחלה ולהבדילה בין מחלות ריאה אחרות לצורך טיפול מתאים בחולים.

בעבודה מסכמת זו נסקרו מספר שיטות מתחום הלמידה העמוקה לסיווג צילומי רנטגן של בית החזה לנבדקים בריאים, נבדקים החולים בקורונה ונבדקים החולים במחלות ריאה אחרות.

דגש מיוחד ניתן בסקירה לרשתות נוירונים המבוססות על קונבולוציה לצד שילובן עם רשתות נוירונים נשנות, שכבות קשב ורשתות נוירונים גנרטיביות. בעבודה בוצע ניתוח השוואתי בין מגוון שיטות הלמידה העמוקה אשר התבסס על מדדי איכות מגוונים כגון דיוק, רגישות, וייחודיות. עמידה במדדי הערכה אלו מהווים קריטריון חשוב לשימוש קליני במערכות סיווג אוטומטיות. במקביל לבחינת מדדי הערכה נבחנו גם זמני ריצה והמשאבים הנדרשים להרצת המודלים המוצעים. לאור ניתוח השוואתי זה נמצא כי ישנן מספר שיטות יעילות לסיווג מחלות ריאה מצילומי רנטגן ובפרט נמצא שמודל מסוג CNN-ATTENTION מביא לתוצאות מדויקות יחסית לגודל קבוצת המדגם שנבחנה ויחסית לכמות המחלקות המסווגות במחקר.

תוכן עניינים

2.....	1. תקציר	2.....
7.....	2. מבוא	7.....
7.....	2.1 תיאור מבנה העבודה	7.....
7.....	2.2 הגדרת הבעיה הרפואית	7.....
7.....	2.2.1 אבחון מחלות ריאה ע"י ניתוח צילומי רנטגן של בית החזה	7.....
7.....	2.2.2 הקורונה והשפעותיה	7.....
8.....	2.2.3 דלקת ריאות (Pneumonia)	8.....
9.....	2.2.4 עכירות ריאה (Lung opacity)	9.....
10.....	2.3 מטרת העבודה וחשיבותה	10.....
12.....	3. למידה עמוקה ככלי סיווג	12.....
12.....	3.1 רשתות נוירונים עמוקות	12.....
12.....	3.1.1 הגדרת רשת נוירונים עמוקה	12.....
14.....	3.1.2 תהליך למידת רשת נוירונים עמוקה	14.....
21.....	3.1.3 אימות הרשת	21.....
22.....	3.1.4 בחינת הרשת	22.....
22.....	3.1.5 בעיית התאמת יתר (Overfitting)	22.....
23.....	3.1.6 כלים למניעת התאמת יתר והגדלת יכולת ההכללה	23.....
25.....	3.2 רשתות קונבולוציה	25.....
26.....	3.2.1 שכבות אופייניות לרשתות קונבולוציה	26.....
31.....	3.2.2 אימון אימות ומבחן הרשת	31.....
32.....	3.2.3 רשתות קונבולוציה הנמצאות בשימוש נרחב	32.....
34.....	3.3 רשתות חוזרות (RNN)	34.....
36.....	3.3.1 תהליך האימון של הרשת	36.....
37.....	3.3.2 אימות ומבחן רשת נשנה	37.....
38.....	3.3.3 ארכיטקטורת הרשת הנשנת	38.....
38.....	3.3.4 בעיית הגרדיאנט הנעלם והמתפוצץ	38.....

39.....	תא LSTM	3.3.5
41.....	מנגנון הקשב וטרנספורמרים	3.4
41.....	תיאור סכמטי מנגנון (שכבת) הקשב	3.4.1
42.....	תיאור פורמלי של מנגנון הקשב	3.4.2
44.....	טרנספורמרים	3.4.3
48.....	רשתות גנרטיביות - GAN	3.5
48.....	מבנה רשת גנרטיבית	3.5.1
49.....	אימון הרשת	3.5.2
51.....	פונקציית ההפסד של GAN	3.5.3
53.....	מבחן והערכת ביצועים ברשתות גנרטיביות	3.5.4
54.....	בעיות בפונקציית ההפסד ברשתות גנרטיביות	3.5.5
56.....	שיטות לסיווג מחלות ריאתיות באמצעות צילומי רנטגן	4.
56.....	סיווג מחלות ריאתיות באמצעות רשתות קונבולוציה	4.1
58.....	מקרה בוחן לסיווג מחלות ריאה באמצעות רשתות קונבולוציה	4.1.1
58.....	מאגר הנתונים	4.1.2
59.....	שיטת המחקר	4.1.3
61.....	תוצאות המחקר ומסקנות	4.1.4
62.....	סיווג מחלות ריאתיות באמצעות רשתות CNN-RNN	4.2
64.....	מקרה בוחן לסיווג באמצעות מודל CNN-LSTM	4.2.1
64.....	מאגר הנתונים	4.2.2
64.....	שיטת המחקר	4.2.3
67.....	תוצאות המחקר ומסקנות	4.2.4
69.....	סיווג מחלות ריאתיות באמצעות רשת קונבולוציה ומנגנון קשב	4.3
70.....	מקרה בוחן לסיווג עם רשת קונבולוציה המשולבת מנגנון הקשב	4.3.1
71.....	מאגר הנתונים	4.3.2
71.....	מהלך המחקר	4.3.3
73.....	תוצאות המחקר ומסקנות	4.3.4

76.....	סיווג מחלות ריאתיות באמצעות רשתות GAN-CNN	4.4
76.....	הקדמה	4.4.1
76.....	מטרות המחקר	4.4.2
77.....	מאגר הנתונים	4.4.3
77.....	מהלך המחקר	4.4.4
82.....	תוצאות המחקר ומסקנות	4.4.5
84.....	מדדי הערכה לניתוח השוואתי	5
85.....	מדד Accuracy	5.1
86.....	מדד Precision	5.2
87.....	מדד Sensitivity	5.3
87.....	מדד Specificity	5.4
88.....	עקומת ROC	5.5
89.....	פונקציית ההפסד LOSS	5.6
91.....	ניתוח השוואתי של השיטות המוצגות	6
94.....	סיכום ומסקנות	7
95.....	המלצות למחקר עתידי	8
96.....	ביבליוגרפיה	9

טבלת איורים:

10.....	איור 1 - צילומי רנטגן של נבדקים [31]
14.....	איור 2 - אילוסטרציה של רשת נוירונים [6]
15.....	איור 3 - סכימת למידה מונחית [8]
18.....	איור 4 - אילוסטרציה של מסלול התנועה של אלגוריתם גרדיאנט דיסנט [19]
21.....	איור 5 - אילוסטרציה של פונקציית אקטיבציה מסוג סיגמואיד [16]
23.....	איור 6 - השוואה בין אפשרויות להתאמת המודל לבסיס הנתונים overfitting, underfitting, balanced [4]
26.....	איור 7 - מבנה רשת קונבולוציה לדוגמה [8]
27.....	איור 8 - פעולת הקונבולוציה [14]
27.....	איור 9 - התקדמות הגרעין על פני תמונת הקלט בפעולת הקונבולוציה; הגרעין מתחיל מהפינה השמאלית העליונה ומתקדם לכיוון הפינה הימנית התחתונה של התמונה [21]
29.....	איור 10 - פעולת הריפוד. משמאל פעולת הקונבולוציה ללא ריפוד. ניתן לראות שגודל תמונת הפלט קטן יותר בהתאם לגודל הגרעין. מימין הקלט המקורי מקבל ערכי ריפוד (מופיעים בקווים מקווקווים דהויים) בכדי שפעולת הקונבולוציה תתקיים גם על דפנות הקלט [14]

29.....	איור 11 - שימוש בגודל צעד משתנה במישור האופקי ובמישור האנכי [8].
12.....	איור 12 - פעולת איגום (Pooling); הפעולה מתבצעת על תמונת הקלט בגודל X55 (מימין) עם גרעין בגודל X33. המטריצה משמאל היא מטריצת הפלט של הפעולה, והתא המסומן באפור בה הוא התוצאה של בחירת ערך המקסימום מאותו אזור המסומן באפור בתמונת הקלט.
30.....	איור 13 - פעולת האיגום. מימין איגום בשיטת AVERAGE ומשמאל איגום בצורת MAX [14].
31.....	איור 14 - הפעלת פעולת האיגום על מישור התמונה [14].
33.....	איור 15 - דוגמה מתוך רשת 101 RESNET [20].
36.....	איור 17 - ארכיטקטורת RNN בסיסית [14].
39.....	איור 18 - ארכיטקטורת LSTM [8].
44.....	איור 19 - מנגנון קשב רב ראשי [14].
46.....	איור 20 - מודל הטרנספורמר [7].
47.....	איור 21 - ייצוג בינארי של מספרים, MSB משתנה בתדירות שונה לעומת LSB [14].
47.....	איור 22 - המחשה לקצב השינוי של פונקציות טריגונומטריות בעלות תדרים משתנים [14].
49.....	איור 23 - ארכיטקטורת מודל GAN [8].
50.....	איור 24 - תהליך למידת המחולל [8].
51.....	איור 25 - תהליך למידת המפלה [8].
59.....	איור 26 - רשתות קונבולוציה [5].
60.....	איור 27 - רשת נוירונים [5].
63.....	איור 28 - שימוש באלגוריתם ROI לזיהוי הריאות [25].
65.....	איור 29 - תיאור סכמטי של תהליך אימון המודלים [31].
66.....	איור 30 - ארכיטקטורת רשת CNN-LSTM [31].
67.....	איור 31 - המודל המלא CNN- LSTM [31].
72.....	איור 32 - תיאור רשת הנוירונים [12].
72.....	איור 33 - כמות הנוירונים לפי בלוקים [12].
74.....	איור 34 - מהירות הריצה של מודלים שונים [12].
75.....	איור 35 - תיאור פונקציית ההפסד ורמת הדיוק בתהליך האימון. [12]
78.....	איור 36 - תהליך אימון הרשת הגנרטיבית [1].
78.....	איור 37 - ארכיטקטורת המחולל [1].
79.....	איור 38 - ארכיטקטורת המפלה [1].
79.....	איור 39 - תאי קידוד ופירוש [1].
80.....	איור 40 - פלט רשת הGAN [1].
81.....	איור 41 - מבנה המערכת בשלב הוצאת הפיצ'רים [1].
81.....	איור 42 - תהליך הוצאת הפיצ'רים ברשת [1].
81.....	איור 43 - וויזואליזציה של תוצרי אלגוריתם BRISK SIFT [1].
88.....	איור 44 - אילוסטרציה של עקומת ROC.
89.....	איור 45 - אילוסטרציה של פונקציית הפסד.

2. מבוא

2.1 תיאור מבנה העבודה

בעבודה זו מוצגת סקירה של גישות נפוצות בלמידה עמוקה לסיווג מחלות ריאה מצילומי רנטגן של בית החזה. בעבודה נסקרו והשוו מספר ארכיטקטורות רשתות נוירונים לסיווג מחלות ריאה כדוגמת קורונה, דלקת ריאות וויראלית ונגיפית, עכירות ריאה ומחלות נוספות. העבודה מחולקת למספר פרקים. בפרק 2 מופיעה סקירה על הבעיה הרפואית קרי מחלות ריאה שונות כגון קורונה, דלקת ריאות ועכירות ריאות. בפרק 3 מוצגת סקירה מקיפה על למידה עמוקה וסוגים שונים של רשתות נוירונים. פרק 4 סוקר ארבע שיטות מרכזיות לסיווג צילומי רנטגן של בית החזה במטרה לזהות ולסווג נכון את מצבם הרפואי של נבדקים החולים במחלות ריאה. פרק 5 עוסק במדדי הערכה לניתוח תוצאות מודלים של למידה עמוקה. בפרק 6 מבוצע ניתוח השוואתי בין השיטות השונות שנסקרו.

2.2 הגדרת הבעיה הרפואית

2.2.1 אבחון מחלות ריאה ע"י ניתוח צילומי רנטגן של בית החזה

צילום רנטגן של בית החזה היא בדיקת הדמיה בטוחה וזולה המצויה בשימוש נרחב שנים רבות. הבדיקה אינה פולשנית וכרוכה בחשיפה למינון נמוך של קרינה מייננת. צילום בית החזה מספק תמונה של איברים פנימיים שבבית החזה כגון הלב, הריאות, דרכי הנשימה, כלי הדם, החוליות ועצמות בית החזה ועל ידי כך יכול לסייע לרופא לאבחן מגוון רחב וגדול של מחלות ריאה שונות כולל דרגת חומרתן. השימוש בכלי זה דורש את אבחנתו של רופא מומחה. ככלל מחלות ריאה שונות משתקפות בצורה שונה בצילום רנטגן של בית החזה [28]. ישנן מספר רב של מחלות ריאה הבאות לידי ביטוי בצילומי רנטגן של בית החזה. המחקרים שנסקרו בעבודה זו מסווגים מחלות ריאה רבות ושונות. בפרק זה אסקור את שלושת המחלות העיקריות המופיעות ברוב המחקרים הנידונים והן עכירות ריאות, דלקת ריאות נגיפית וויראלית וקורונה.

2.2.2 הקורונה והשפעותיה

קורונה (COVID) היא משפחת וירוסים הגורמים לטווח רחב של מחלות – החל מצינון וכלה במחלות נשימתיות חמורות, כמו וירוס מרס (MERS-CoV) ומחלת הסארס. כל משפחת נגיפי הקורונה יכולים לעבור בין חיות לבני אדם (מחלות זואוונטיות). למעשה, קיימים זנים נוספים

ממשפחת הקורונה שקיימים בבעלי חיים אך עדיין לא הדביקו בני אדם. COVID-19 הינו זן חדש שנצפה לראשונה בבני אדם והתפרץ לראשונה בסין בשנת 2019 [29].

ככל ששיעור הבדיקות לזיהוי קורונה יורד ברחבי העולם, קשה יותר לדעת כמה אנשים חולים בקורונה ואינם מצהירים על כך בפני גורם רפואי. בתחילת המגיפה, 15% מהאנשים שנדבקו בנגיף נחשבו כאנשים הזקוקים לטיפול רפואי בבתי חולים ולחמצן. הערכות עדכניות יותר מצביעות על כך שנדרש אשפוז בכ-3% מהאנשים החולים בקורונה. השינוי בכמות האנשים המבקשים טיפול רפואי נובע בחלקו משימוש נרחב בקרב האוכלוסייה העולמית בחיסונים, בחלקו בשל שינויים בנגיף (בעיקר גרסאות האומיקרון), ובחלקו בשל הזמינות של טיפולים רפואיים ממוקדים הנמצאים בשימוש ברחבי העולם [29].

אוכלוסיות בסיכון כגון קשישים או אנשים עם מחלות רקע זקוקים ביתר שאת לאבחון מוקדם ולטיפול מתאים בבית חולים. במצבים נדירים, ילדים יכולים לפתח תסמונת דלקתית חמורה מספר שבועות לאחר ההדבקה ולכן הם זקוקים לאבחון מוקדם ולטיפול ראוי. יש לציין שרוב האנשים מחלימים באופן מלא מבלי להזדקק לטיפול בבית חולים.

בדומה למחלות נשימתיות אחרות קורונה יכולה לגרום לתסמינים שונים ברמות חומרה שונות. התסמינים העיקריים הם נזלת, כאבי גרון, שיעול וחום, אך ישנם גם תסמינים חמורים ונדירים יותר כגון דלקת ריאות, קשיי נשימה וקוצר נשימה, כאבי ראש וסחרחורות, כאבי שרירים, נמלולים בגפים ועוד. במקרים חריגים ההדבקה עלולה להיות מסכנת חיים ולהוביל לכשל כלייתי ולמוות. ארגון הבריאות העולמי מציין כי אנשים מבוגרים מעל גיל 60 ובעלי מחלות רקע כגון אנשים בעלי בעיות בלחץ הדם, אנשים סכרתיים ואנשים בעלי מחלות כרוניות נוספות מצויים בסכנה ממשית ומוגברת מהמחלה [29].

לקורונה ישנם גם השפעות ארוכות טווח על המטופלים שמתועדות בספרות המקצועית ולעיתים קרובות מכנים את השפעות אלו בשם LONG COVID. הימנעות מהמחלה כוללת מספר כלים בסיסיים כגון שמירה על מרחק חברתי, חבישת מסכה בחללים צפופים או חללים המאווררים בצורה גרועה, שמירה על היגיינת ידיים, כולל אימוץ כללי התנהגות נשימתיים (כיסוי הפה והאף במרפק כפוף או רקמה בזמן שיעול או עיטוש), נטילת ידיים קרובות וחיטוי משטחים משותפים וכמובן התחסנות חברתית [29].

2.2.3 דלקת ריאות (Pneumonia)

דלקת ריאות היא זיהום של שקיקי הריאה שהם שקיקי אוויר זעירים שנמצאים בתוך הריאות הנקראים גם נאדיות. בשקיקים אלו נקלט החמצן מהאוויר בתהליך הנשימה ובהם מועבר

חמצן לזרם הדם. זיהום או פגיעה בשקיקים אלו עלול לגרום לאספקה לא סדירה של חמצן לכלי הדם.

דלקת ריאות יכולה להיגרם כתוצאה מזיהום חיידקי או נגיפי ולעיתים נדירות גם מפטריות. לעיתים מחלה נגיפית כמו שפעת עלולה לגרום לדלקת ריאות. תסמיניה של דלקת ריאות הם רבים וכוללים חום, שיעול, קוצר נשימה, כאבים בחזה וקצב נשימות מוגבר. מידת ההדבקה של דלקת ריאות תלויה בגורם למחלה. דלקת ריאות הנגרמת מנגיף הקורונה מדבקת יותר מדלקת ריאות הנגרמת מחיידק הפנימוקוק. יחד עם זאת על פי רוב אין צורך לבודד את החולה בכדי למנוע הדבקה. אבחון דלקת ריאות מבוצע על ידי תשאול החולה אחר התסמינים שתוארו לעיל, או על ידי ממצאים מבדיקה גופנית או האזנה לריאות [9].

דלקת ריאות ניתנת לאבחנה בתמונת רנטגן של בית החזה ומופיעה בצורת אזורים אטומים בריאה שמשקפים בועיות ריאה המלאות בנוזלים דלקתיים. בפועל קשה לאבחן האם דלקת ריאות היא חיידקית או נגיפית על בסיס תשאול בלבד וגם התבוננות בתמונת רנטגן איננה נותנת פתרון וודאי.

עקב הקושי בזיהוי ואבחון המחלה הטיפול לרוב מתבסס על אנטיביוטיקה הניתנת ישירות לווריד או דרך הפה. משך הטיפול במחלה הולך ומתקצר בשנים האחרונות במסגרת המגמה העולמית להפחית ככל האפשר את השימוש באנטיביוטיקה כדי לנסות ולהימנע מתופעות לוואי אצל המטופל ויצירת זנים עמידים לאנטיביוטיקה [9].

2.2.4 עכירות ריאה (Lung opacity)

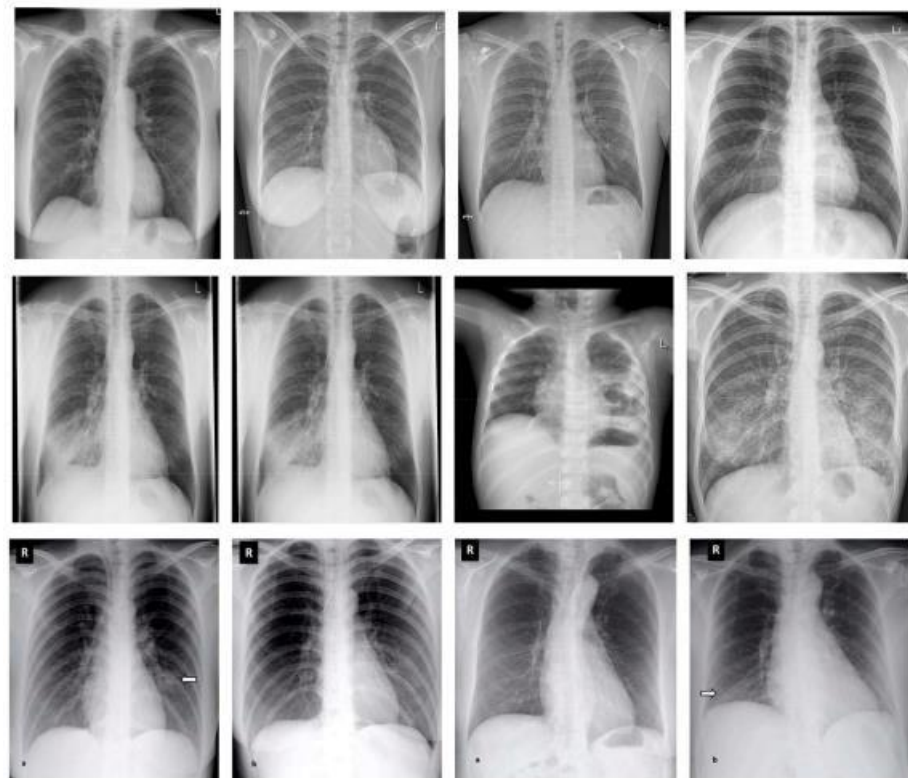
עכירות ריאה היא מושג רדיולוגי המתייחס לאזורים אפורים או מעורפלים בתמונות שנעשו על ידי סריקות CT או צילומי רנטגן של הריאות. עכירות מצביע על צפיפות מוגברת באזורים בהם היא מופיעה. ריאות בריאות נראות בצבע שחור בבדיקת CT או בצילום רנטגן. הצבע השחור האחד מעיד שהריאות נקיות מחסימות או הצטברויות של נוזלים. כאשר אזורים אפורים נראים במקום זאת, הדבר מצביע על כך שמהו ממלא חלקית את האזור בתוך הריאות כגון נוזל או מוגלה [23].

עכירות ריאה יכולה להיות תסמין זמני ותוצאה של מחלה קצרת טווח. במקרים אחרים, העכירות יכולה להעיד על מצב כרוני או חמור יותר כגון זיהום או תהליך דלקתי אחר. העכירות יכולה לנבוע ממגוון של מחלות כרוניות כגון דלקת ריאות, קורונה, סרטן ריאה ומחלות נוספות. יחד עם זאת עכירות יכולה להופיע גם במצבים בהם המטופל מעשן בצורה קבועה.

העכירות יכולה להופיע במגוון דרכים בתמונת הרנטגן, בהתאם למחלה ותנאי הסביבה. העכירות יכולה להיות מפוזרת בחלל הריאה על פני אונות שונות או בצורות פסיפס שונות בדרגות עכירות משתנות. הטיפול בעכירות הריאה משתנה בהתאם לאבחון הרופא ונע בין מתן סטרואידים, אנטיביוטיקה, חמצן ולעיתים בחדר הניתוחים [23].

2.3 מטרת העבודה וחשיבותה

סיווג ידני של מחלות ריאה באמצעות צילומי רנטגן של בית החזה היא משימה מורכבת. נדרש ניסיון רב בכדי להבחין ולהבדיל בין סוגים שונים של מחלות ריאה באמצעות צילומי רנטגן. באיור 1 מודגמים הבדלים בין צילומי רנטגן של אנשים בריאים וחולים כאשר שורת התמונות הראשונה מכילה תמונות של חולי קורונה, השורה השנייה מכילה תמונות של חולי דלקת ריאות כרונית ובשורה השלישית אנשים בריאים [31].



איור 1 - צילומי רנטגן של נבדקים [31].

ישנם מודלים שונים המבוססים על טכניקות שונות של למידה עמוקה המנסים לסווג מחלות ריאה בצילומי רנטגן. חלק נרחב מהם מבוססים על מודלים הבנויים מרשתות קונבולוציה כדוגמת VGG16 או רשתות קונבולוציה מוכרות אחרות המבצעים תהליך סיווג יעיל ומהיר יותר מתהליך הסיווג האנושי. לצד הגידול הרב בכמות המחקרים שבוצעו בנושא הלך וגדל גם היקף המידע המתועד בנושא ובמיוחד עבור מחלת הקורונה ומאגרי המידע המגוונים

כדוגמת צילומי רנטגן או CT של חולי קורונה הלכו וגדלו אף הם. נמצא, שככל שמאגרי המידע גדלו, כך חלה ירידה ברמת דיוק סיווג מחלת קורונה באמצעות רשתות קונבולוציוניות [1].

עובדה זו ניתנת להסבר בכמה דרכים. ראשית הגידול במאגרי המידע נוצר על ידי שימוש במכונות צילום רנטגן באיכויות שונות ובמקומות שונים ברחבי העולם. הדאטה שנאסף במאגרי המידע השונים מגוון מגיע ממקורות שונים ובאיכות משתנה. שנית כמות התמונות וכמות המחלות המתועדות גדל אף הוא. המחקרים הראשונים בתחום סקרו מספר קטן של מחלות ריאה (בין 2-3 סוגי מחלות) ומחקרים מאוחרים סוקרים כמות גדולה יותר של מחלות (בין 5-6 סוגי מחלות שונים).

שימוש בבינה מלאכותית ככלל ובטכניקות של למידה עמוקה בפרט מאפשרים לבצע ניתוח בכלים נוספים מעבר לסל הכלים הקיימים עבור רופאים ברמות מומחיות שונות בצורה יעילה וקלה יותר. בעוד שבחינה אנושית של צילומי רנטגן של בית הריאה ע"י רופא דורשת מומחיות, שימוש בכלים של למידה עמוקה מאפשרים לבצע סיווג בקלות ובנגישות רבה יותר [17].

שימוש בסיווג אוטומטי מבוססת AI של תמונות רנטגן בית חזה, כפי שאציג בעבודה זו, עשוי להוות כלי עזר רפואי ולסייע לרופאים לאבחן ולזהות מחלות ריאה מסוגים שונים. מטרת העבודה היא לסקור את התחום המחקרי בנושא, לעמוד על ההבדלים בין שיטות הסיווג השונות ולהצביע על הכלים היעילים ביותר לסיווג מרחב מחלות גדול ביותר באמצעות כלי למידה עמוקה.

3. למידה עמוקה ככלי סיווג

3.1 רשתות נוירונים עמוקות

למידה עמוקה היא תת קבוצה של תחום למידת המכונה. בליבה של טכנולוגיה זו נמצאת "הרשת העצבית העמוקה". רשת עצבית היא מודל חישובי אשר נבנה בהשראת המוח האנושי ומכיל שכבות מחוברות של יחידות חישוביות קטנות הנקראות נוירונים. רשתות נוירונים הן אבן הבסיס של תחום הלמידה העמוקה. יישומיה השונים של הלמידה העמוקה נסובים סביב יצירת דפוסים שונים של רשתות נוירונים כגון רשתות מחוברות בצורה מלאה (Fully connected), רשתות המבוססות על פעולת הקונבולוציה, רשתות בעלות תאים נשנים, רשתות גנרטיביות ומנגנונים נוספים.

מלבד מחקר ופיתוח של סוגי רשתות שונים וחיבורים שונים בין נוירונים, קיים מחקר רב על התהליכים הנדרשים לטובת שיפור הכוח החישובי של רשתות נוירונים עמוקות וביצוע אופטימיזציה לתהליך הלימוד של רשתות אלו. ישנם כלים רבים בתחום הלמידה העמוקה שמטרתם להביא לאופטימיזציה מרבית של רשתות הנוירונים כגון פונקציות אופטימיזציה, פונקציות הפסד, פונקציות אקטיבציה וכלי נרמול [8].

כיום, אנו עדים לשימוש נרחב ביכולות הלמידה העמוקה במגוון תחומים ויישומים. החל מעיבוד מידע רפואי וסיוע באבחון מחלות כנדון בעבודה זו וכלה בתעשיות רבות אחרות שבהן יש צורך להסתייע ביכולות ניתוח נתונים מתקדמות [8].

3.1.1 הגדרת רשת נוירונים עמוקה

ניתן להגדיר את רשת הנוירונים באמצעות מודל של גרף מכוון ללא מעגלים בעל פונקציית משקל על פני הקשתות. הצמתים בגרף מייצגים נוירונים או יחידות, והקשתות מייצגות חיבורים בין נוירונים. בגרף מסוג זה המידע זורם דרך צמתי הקלט ועד לצמתי הפלט מבלי לעבור במעגל. כל צומת בגרף למעט צומתי הקלט והפלט מקבל מידע מצומת קלט אחת או יותר ושולח את המידע שחושב בתוכו לצומת אחת או יותר.

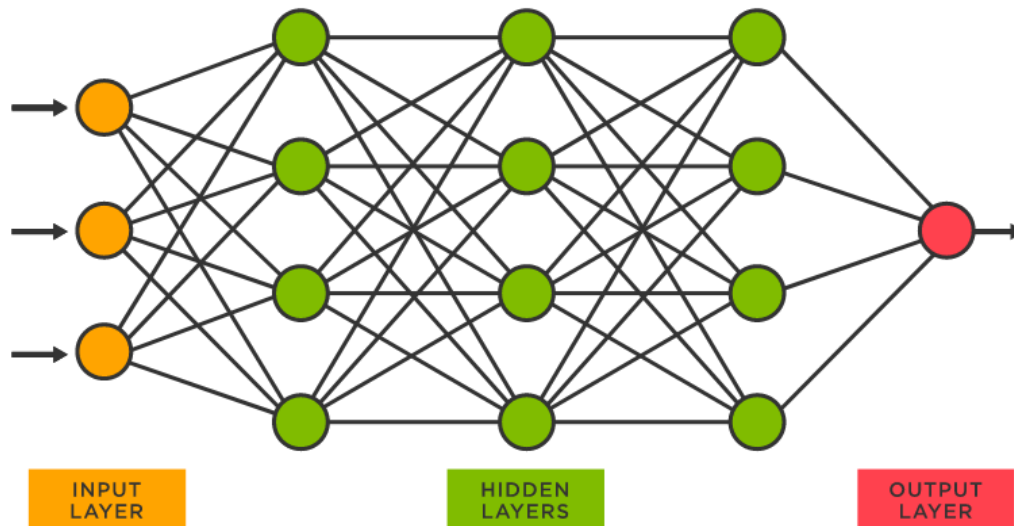
כל קשת בגרף משויכת למשקל. המשקל בין זוג נוירונים מייצג את הקשר ביניהם. פונקציית המשקל מקצה לכל קשת ערך מספרי המייצג את השפעת הנוירון במקור הקשת על הנוירון ביעד הקשת.

רשת נוירונים עמוקה (DNN) מאופיינת במספר רב של שכבות נסתרות בין שכבות הקלט והפלט. כל שכבה מורכבת מצמתים מחוברים, המבצעים חישובים על נתוני הקלט. ב-DNN, המידע זורם דרך השכבות הללו משכבת הקלט, בה מוזנים הנתונים הגולמיים, דרך שכבות

נסתרות, שבהן מתרחשות טרנספורמציות וחישובים שונים ועד לשכבת הפלט המייצרת את הפלט הסופי התלוי בהגדרת הרשת קרי הפלט יכול להיות ערך בודד בבעיית רגרסיה או ערך ההסתברות עבור סיווג הקלט למחלקה נתונה מתוך סל המחלקות הקיימות בתהליך האימון של הרשת. המונח "עמוק" ברשתות נוירונים מתייחס לנוכחותן של שכבות נסתרות מרובות [14].

רשת נוירונים מורכבת כאמור משכבות המחוברות יחדיו. ניתן לחלק את סוגי השכבות ברשת הנוירונים לשלוש סוגים עיקריים: שכבת קלט, שכבות נסתרות ושכבת פלט:

- שכבת הקלט (Input Layer) – שכבה זו אחראית על קבלת הקלט אשר מתקבל ממקור חיצוני. תמיד ישנה שכבת קלט אחת לרשת נוירונים. שכבת הקלט מקבלת את נתוני הקלט שמוכנסים לרשת הנוירונים, מבצעת את חישוביה דרך הנוירונים שלה והפלט של שכבה זו מועבר לשכבה הבאה ברשת, בין אם מדובר ישירות בשכבת הפלט או בשכבה נסתרת המחוברת את שכבת הקלט לשכבת הפלט.
- שכבה נסתרת (Hidden Layer) – שכבות נסתרות נמצאות בין שכבת הקלט לשכבת הפלט, כאשר בכל רשת נוירונים עמוקה יש לפחות שכבה נסתרת אחת או יותר. שכבות אלה מאפשרות לרשת הנוירונים ללמוד דפוסים מורכבים מנתוני הקלט לרשת ומשפרים את העוצמה החישובית של רשת הנוירונים. מספר שכבות הקלט איננו חסום מלמעלה ויכולות להיות מספר רב של שכבות נסתרות ברשת נוירונים עמוקה.
- שכבת הפלט (Output Layer) – שכבת הפלט אחראית על חישוב הפלט הסופי של הרשת. ברשת נוירונים תמיד ישנה שכבת פלט אחת, אשר כוללת מספר נוירונים כמספר האפשרויות לתיוג מחלקות הבעיה. באם מדובר בבעיית רגרסיה שבמסגרתה נדרשת הרשת לצפות ערך בודד, בשכבת הפלט יהיה נוירון בודד. שכבת הפלט מקבלת קלט מהשכבה הקודמת לה, מבצעת את חישוביה בעזרת הנוירונים (או נוירון בודד בבעיית רגרסיה) שלה ומחזירה את הפלט הסופי של הרשת.



איור 2 - אילוסטרציה של רשת נוירונים [6]

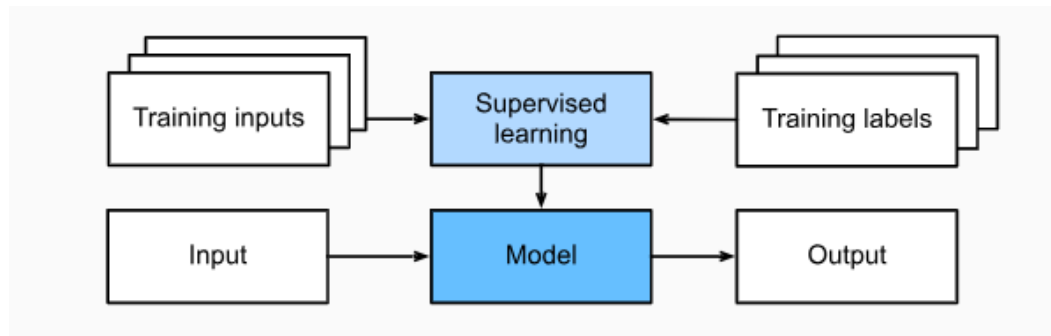
באיור 2 ניתן לראות אילוסטרציה של רשת נוירונים עמוקה. באיור זה כל נוירון בשכבות הנסתרות מחובר לכל הנוירונים בשכבה הקודמת לו, כך שהקשתות הנכנסות אליו ממושקלות לפי פונקציית המשקל w . ניתן לראות כי באיור קיימות 3 שכבות נסתרות הצבועות בירוק ובכל שכבה ארבע נוירונים (אין הכרח כי מספר הנוירונים בכל השכבות החבויות יהיה זהה). במודל שכבת קלט בודדת הצבועה בכתום ובה שלושה נוירונים האמונים על קבלת הקלט, ושכבת פלט בודדת הצבועה בבורדו בעלת נוירון בודד. הנוירונים מחוברים יחדיו באמצעות סדרה של משקולות. בתהליך הלימוד של הרשת משקולות אלו ישנו את ערכם בצורה איטרטיבית בהתאם לערכי פונקציית ההפסד ואלגוריתם האופטימיזציה עליהן נדון בהמשך. ככלל ניתן לומר שגודל הרשת משפיע לטובה על ביצועיה עד גבול מסוים. הגדלת רשת הנוירונים במספר רב של שכבות נסתרות ונוירונים רבים בכל שכבה איננו מבטיח בצורה חד משמעית כי ביצועי הרשת, על משימותיה השונות, ישתפרו בוודאות [8].

3.1.2 תהליך למידת רשת נוירונים עמוקה

בעבודה זו אנו מתמקדים בבעיות המשתייכות לתחום הלמידה המונחית (Supervised Learning) בה הלמידה מתבצעת מקבוצת דוגמאות מתויגות.

בלמידה מונחית אנו משתמשים במאגר מידע הכולל דוגמאות ותיוגים עבור כל דוגמה. בפרט בעבודה זו, מאגר הדוגמאות הינו תמונות רנטגן בעלות תיוגים שבוצעו בצורה ידנית בין אנשים בריאים ואנשים החולים במחלות ריאה שונות.

כאשר אנו מאמנים רשת נוירונים אנו מתכוונים כי מטרתנו היא ליצור מודל אשר בהינתן דוגמת קלט יוכל לחזות את התיוג שלה או לחזות את ערכה. בנוסף מטרתנו היא כי רמת הוודאות של המודל תהיה גבוהה ככל הניתן ושיעור הטעויות של המודל על גבי מרחב הדוגמאות יהיה קטן ככל הניתן. סוג למידה זה מכונה "מונחה" או "מפוקח" כיוון שאנו מנחים את המודל לגבי התיוג הנכון עבור כל אחת מדוגמאות הקלט שבאמצעותו מתבצע אימון המודל [8].



איור 3 – סכימת למידה מונחית [8].

באיור מעלה מוצג תהליך הלמידה המונחה של רשת הנוירונים. אימון רשת נוירונים באמצעות מערך נתונים מתיוג הנקרא גם קבוצת האימון (Training Set). במהלך תהליך האימון המודל מתאים ומשנה את ערכי המשקולות שבין הנוירונים מהם הוא מורכב כדי למזער את השגיאה בין תיוג המודל ובין תיוג האמת שהתקבלו באמצעות הלמידה המונחית. לאחר סיום אימון המודל, באפשרותנו להזין למודל דוגמה לא מתיוגת ולקבל כפלט את התיוג של רשת הנוירונים עבורה.

תהליך האימון מתבצע באמצעות אופטימיזציה של ערכי פונקציית ההפסד (המוגדרת מראש) אשר מודדת ומכמתת את הפער בין הפלט החזוי של המודל ובין הפלט האמיתי של הדוגמה המוזנת אליו.

מלבד קבוצת האימון נעשה שימוש בקבוצות מידע נוספות הנקראות קבוצת האימות (Validation Set) וקבוצת המבחן (Test Set) עליהן ארחיב בסעיפים הבאים.

לפני שנדון בפונקציית ההפסד ואלגוריתם האופטימיזציה של רשת הנוירונים נציג את תהליך האימון השלם של רשת הנוירונים.

כאמור, ניתן להגדיר את המודל של רשת נוירונים עמוקה כדלהלן:

$$f(x_i; W): X \rightarrow L$$

הרשת מקבלת כקלט דוגמאות מהקבוצה X , משתמשת בפונקציית המשקל על פני הקשתות W , ומחזירה תיוג $l \in L$ כלשהו, מבין המחלקות הקיימות בבעיה הנתונה. מטרת תהליך האימון הינה להביא לדיוק טוב ככול הניתן של התיוג של המודל על ידי מזעור ערכה של פונקציית ההפסד שהוגדרה מראש על ידי תהליך אופטימיזציה.

אימון רשת הנירונים כולל את השלבים המרכזיים הבאים:

1. אתחול משקולות - משקלי הרשת מאותחלים לרוב באופן אקראי או באמצעות יוריסטיקה מוגדרת מראש. במהלך האימון, משקלים אלו יותאמו כדי למזער את ההבדל בין התיוג של המודל ובין ערך האמת של כל דוגמה המוזרמת למודל. קיימות מספר שיטות לקביעת ערכי הבסיס של משקולות המודל טרום שלב האימון כגון בחירת משקולות בהתפלגות מסוימת או בטווח ערכים מוגדר.
2. הפצה קדימה או פעפוע לפנים - נתוני הקלט מוזנים לרשת ו"מפועפעים" קדימה בין שכבות הרשת. הסכום המשוקלל של כלל חישובי המשקולות בנירונים השונים על פני השכבות השונות מועברים קדימה ומסוכמים לתוצאה סופית אשר מסמלת את הפלט של הרשת. ערך הסיווג של הרשת בין אם מדובר בהסתברות הסיווג מבין מרחב מחלקות קבוצת המבחן ובין אם מדובר בערך בודד עבור בעיית רגרסיה הוא ערך מספרי.
3. חישוב הפסד - ההפרש בין הפלט של הרשת לבין סיווג האמת של הדוגמה שהוכנסה למודל מכומת ונמדד באמצעות פונקציית הפסד. פונקציה זו מודדת את ביצועי הרשת, ומספקת ערך מספרי המייצג את גודל השגיאה בין ניבוי המודל ובין ערך האמת. קיימות סוגים שונים של פונקציות הפסד בהתאם למישור הבעיה שהרשת אמורה לפתור.
4. פעפוע לאחור – הרשת משתמשת באלגוריתם אופטימיזציה כדי למזער את ההפסד שחושב ע"י פונקציית ההפסד. הפעפוע לאחור כרוך בחישוב שיפוע ההפסד ביחס למשקולות השונות בהתאם למיקומן על גבי רשת הנירונים. מידע זה משמש כדי להתאים את המשקולות בכיוון שמפחית את גודל השגיאה במטרה למזער אותה ככול שניתן.
5. עדכון משקלים – משקולות הרשת מתעדכנות באופן איטרטיבי על סמך השיפועים המחושבים של אלגוריתם האופטימיזציה וגודל השגיאה של פונקציית ההפסד. תהליך זה חוזר על עצמו על פני מספר איטרציות או אפוקים, עד שהרשת מתכנסת למצב שבו ההפסד ממוזער בנקודת אופטימום.

תהליך האימון תלוי בערכיהן של שני כלים מרכזיים – פונקציית ההפסד ואלגוריתם האופטימיזציה. על אף שקיימות סוגים רבים של פונקציות הפסד ואלגוריתמי אופטימיזציה

אתמקד בפרק זה בפונקציית הפסד מסוג אנטרופיה צולבת קטגוריאליה ובאלגוריתם אופטימיזציה מסוג מורד הגרדיאנט (GD).

פונקציית ההפסד הינה פונקציה המכמתת את הפער בין ערך החיזוי של המודל ובין ערך התיאור האמיתי של דוגמה ספציפית שהוזנה לרשת. מטרת תהליך האימון של הרשת היא למזער את גודל השגיאה של פונקציית ההפסד ובכך להביא את המודל לביצוע תחזיות מדויקות. אציין כי פונקציית ההפסד וערכה יכול לשמש אותנו גם בתהליך הערכת ביצועי הרשת ולא רק בשביל תהליך אימון הרשת, אך אדון בפרט זה בהמשך.

בעוד שישנם סוגים שונים של פונקציות הפסד, פונקציית האנטרופיה הצולבת הקטגוריאליה (CCE) היא פונקציה נפוצה המשמשת למשימות סיווג רב מימדיות. הפונקציה מודדת את השוני בין התפלגות ההסתברות החזויה של המודל לבין ההתפלגות האמיתית של תיאור המחלקה אליה שייכת הדוגמה. במילים אחרות, הפונקציה מודדת את גודל הטעות בין מה שהמודל תיג בפועל ובין תיאור האמת של דוגמה נתונה [14].

נניח כי קיימות N מחלקות במרחב הדוגמאות מעוניינים לסווג לאחר הזנת דוגמה לרשת הנורונים המודל מחשב את התפלגות החיזוי של אוסף המחלקות עבור הדוגמה שהוזנה לו, בדרך כלל על ידי שימוש בפונקציית אקטיבציה כגון סיגמאנויד, RELU או Tanh. עבור כל דוגמה שהוזנה לרשת, יש בידינו את התיאור האמיתי שלה. התפלגות התיאורים של הרשת מקודדת לרוב בצורת one-hot כלומר ווקטור המורכב מערכי אפס ואחד, אשר האינדקס של המחלקה המתאימה לדוגמה הינו אחד ושאר איברי הווקטור הם אפס.

עבור כל דוגמה ניתן לחשב את רמת האנטרופיה הצולבת בעזרת הנוסחה הבאה:

$$CCE = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log y_{ij}^{\wedge}$$

כאשר:

N הוא כאמור מספר הדוגמאות בסט הנתונים.

C הוא מספר המחלקות בבעיית הסיווג.

y_{ij} הוא ההתפלגות האמיתית של תיאור המחלקה (ווקטור בעל הערכים אפס ואחד).

$y_{ij}^{\wedge}$ היא ההתפלגות החזויה שעבורה הדוגמה ה*i* שייכת למחלקה ה*j* כפי שנחזה על ידי המודל.

כיוון שישנו רק תיאור אמת אחד לכל דוגמה, כלל ערכי המשוואה מתאפסים למעט ההפרש בין ערך החיזוי של המודל (עליו מבוצעת פעולת הלוגריתם) לסיווג האמת של הדוגמה. אוסף

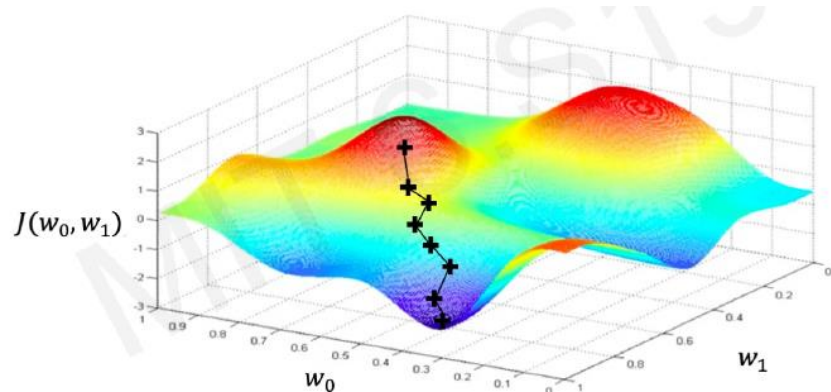
הטעויות נסכם לגבי כלל הדוגמאות שהוזנו לרשת וכך ניתן לחשב את ערך פונקציית ההפסד הכולל על כלל סט הנתונים.

במילים אחרות, הזנת הדוגמה לרשת הנירונים מייצרת פלט הסתברויות או חיזויים של המודל עבור מרחב התיגים האפשריים. ככול שרמת הביטחון של המודל, המבוטא בעוצמת הסתברות שנתן המודל למחלקת הסיווג האמיתית של דוגמה נתונה גדולה יותר כך גודל השגיאה שמחשבת פונקציית ההפסד קטן. ככול שרמת הביטחון של המודל קטנה עבור תיג האמת של הדוגמה, כך גודל השגיאה גדל.

משאנו יודעים לחשב את גודל שגיאת החיזוי של המודל עבור כל סט הנתונים, נותר לדון כיצד ניתן להשתמש בנתון זה בכדי לתקן את המשקולות הפנימיות של רשת הנירונים בכדי להביא לאופטימיזציה של הרשת.

לטובת העניין ניתן להשתמש באלגוריתם אופטימיזציה. מכיוון שפונקציית המשקלים של רשת הנירונים עשויה להיות מורכבת מאוד, אין אלגוריתם דטרמיניסטי יעיל שמוצא את הערך שיביא אותה למינימום הטעות. קיימים מספר אלגוריתמים לביצוע אופטימיזציה לרשתות ניורונים. בעבודה זו אסקור אלגוריתם איטרטיבי בשם Gradient Descend.

העקרון מאחורי האלגוריתם הינו שניתן להביא את ערכה של פונקציית ההפסד למינימום במידה ונעים מנקודת הראשית שעל גרף הפונקציה. בכל איטרציה "נעים" בכיוון ההופכי לכיוון הגרדיינט של הפונקציה באותה נקודה. מזכיר כי הגרדיינט תמיד מצביע בכיוון העלייה החדה ביותר של גרף הפונקציה, ולכן "התנועה" בכיוון ההופכי לו תהיה בהכרח בכיוון הירידה החדה ביותר.



איור 4 - אילוסטרציה של מסלול התנועה של אלגוריתם גרדיאנט דיסנט [19].

אלגוריתם Gradient Descent מוגדר באופן הבא:

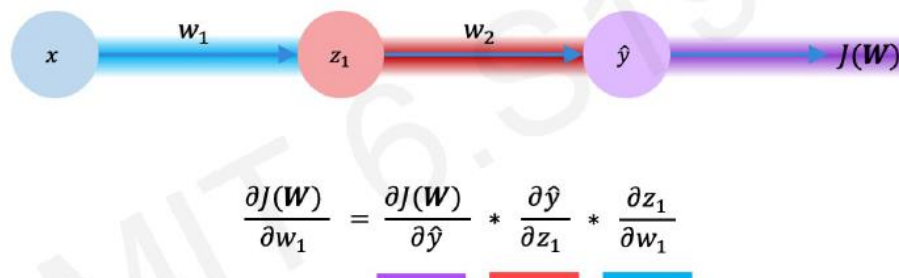
1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Compute gradient, $\frac{\partial J(W)}{\partial W}$
4. Update weights, $W \leftarrow W - \eta \frac{\partial J(W)}{\partial W}$
5. Return weights

אלגוריתם 1 - פסודו קוד גרדיאנט דיסנט [4].

ראשית בכדי להשתמש באלגוריתם עלינו לחשב את הנגזרת החלקית של כל אחת ממשקולות המודל. התהליך מתחיל מחישוב כל הנגזרות החלקיות של משקולות המודל

$$\frac{\partial J(W)}{\partial w_i} \text{ לכל } w_i \in W.$$

איור 6 מדגים את צורת החישוב של נגזרת חלקית אחת עבור המשקולות w_1 . תחילה מחושבת הנגזרת של פונקציית ההפסד שחושבה בסוף הליך הפעפוע לפני (המתואר באיור כ $\hat{y}$). לאחר מכן, בעזרת כלל השרשרת, ניתן לחשב לאחר את כל הנגזרות החלקיות של כלל משקולות הרשת, בעזרת כל הנגזרות החלקיות, ניתן לחשב את הגרדיינט כולו $\frac{\partial J(W)}{\partial W}$.



איור 5 - חישוב הנגזרות החלקיות של משקולות המודל [4].

מרגע שחושב הגרדיינט עבור כל משקולות במודל באמצעות כלל השרשרת ניתן לעדכן את פונקציית המשקלים בכיוון ההופכי לכיוון הגרדיאנט. תהליך זה מתקיים עד שהגרדיינט מתאפס, כלומר פונקציית ההפסד מגיעה לנקודת מינימום. ובשלב זה מסתיים תהליך האימון של הרשת עם החזרת פונקציית המשקל שנמצאה, W^* .

ניתן לתאר את תהליך עדכון המשקולות באמצעות הנוסחה הבאה:

$$\partial_{t+1} = \partial_t - \alpha \nabla J(\partial_t)$$

משוואה 1 - עדכון משקולות גרדיאנט [4].

כאשר ∂_t הוא פרמטר המשקולת באיטרציה t .
אלפא היא קבוע למידה המוגדר מראש.
ו $\nabla J(\partial_t)$ הוא גרדיאנט פונקציית ההפסד ביחס למשקולת מסויימת.

אלגוריתם Gradient Descent אומנם נפוץ בשימוש אך במקביל אליו מקובל לשימוש אלגוריתם נוסף השואב את השראתו מאלגוריתם מורד הגרדיאנט ונקרא הינו Stochastic Gradient Descent, או בקיצור SGD. האלגוריתם מתואר באיור הבא:

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Pick batch of B data points
4. Compute gradient, $\frac{\partial J(W)}{\partial W} = \frac{1}{B} \sum_{k=1}^B \frac{\partial J_k(W)}{\partial W}$
5. Update weights, $W \leftarrow W - \eta \frac{\partial J(W)}{\partial W}$
6. Return weights

אלגוריתם 2-פסודו קוד לאלגוריתם מורד הגרדיאנט הסטוכסטי [4].

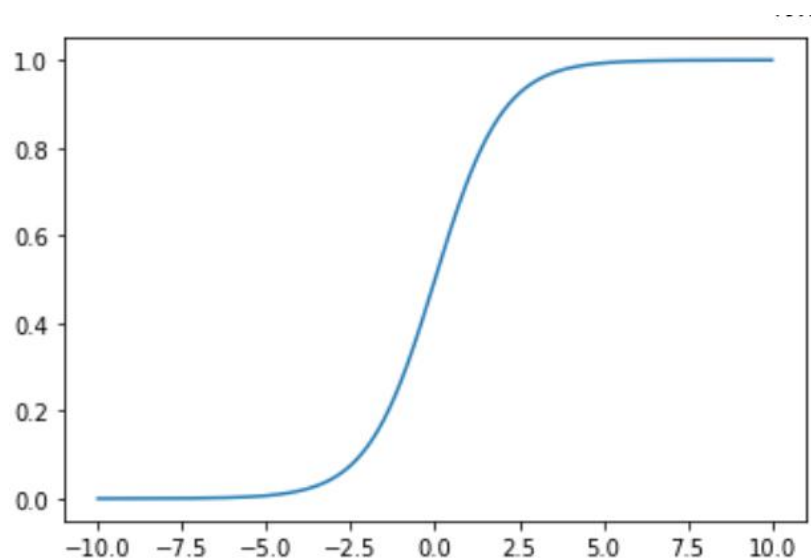
נשים לב כי האלגוריתם דומה מאוד לאלגוריתם Gradient Descent הבסיסי שהוצג קודם לכן עם שינוי אחד עיקרי, והוא שחישוב הגרדיינט ועדכון המשקלים מתבצע על פני קבוצות, Batch-ים, של דוגמאות מקבוצת האימון. השיקול העיקרי לשימוש בשיטה זו הינה היעילות החישובית בחישוב הגרדיינט עבור מספר מסוים של נקודות מאשר עבור כל קבוצת האימון [4].

כלי נוסף הנמצא בשימוש נרחב בעולם הלמידה העמוקה הוא השימוש בפונקציות אקטיבציה. פונקציות אקטיבציה הן רכיבים חיוניים המאפשרים לרשתות הניורונים ליצור מודלים ולתאר תלויות של יחסים מורכבים ולא ליניאריים. פונקציות האקטיבציה הן חלק אינטגרלי מפעולתו של כל אחד מהניורונים במעמקי הרשת. ללא פונקציות אלו, רשתות ניורונים יצטמצמו למעשה למודלים ליניאריים, מה שמגביל את יכולתן ללמוד יחסים מורכבים מתוך הנתונים [8].

פונקציות אקטיבציה פועלות ברמת הנירון הבודד. פונקציות האקטיבציה אינן עוברות תהליך של שינוי והתאמה במהלך תהליך הלימוד כמו משקולות הקשתות המחוברות ניורונים וערכן של פונקציות אלו קבוע לאורך האימון. ישנם מספר סוגים נפוצים של פונקציות אלו:

1. סיגמואיד – פונקציה זו מצמצמת את הפלט של נירון בודד לטווח שבין 0 ל-1. היא משמשת לעתים קרובות בשכבת הפלט של מודלים של סיווג בינארי, כאשר המטרה היא לייצר פלט דמוי הסתברות. פונקציה זו נפוצה בשימוש בעיקר בשכבת הסיווג

- הסופית של רשת הניורונים. הסיבה לכך היא העובדה שהפונקציה מייצרת פלט בטווח סגור שניתן להמרה לאחוזים, כלומר הערך המספרי בין אפס לאחד ניתן להמרה לערך מספרי בין אפס למאה אחוז בקלות. תכונה זו מאפשרת נוחות רבה בקביעת רמת הביטחון הסופית של פלט רשת ניורונים במשימות סיווג.
2. פונקציית טנג'נט היפרבולי ($\tanh$) - בדומה לפונקציה הסיגמואידית, הפונקציה $\tanh$ ממפה את הפלט לטווח שבין -1 ל-1. היא משמשת לעתים קרובות בשכבות נסתרות ברשת הניורונים, ומספקת פלט הנע מסביב לערך אפס, שעוזר להפחית בעיות הקשורות לשיפועים נעלמים או מתפוצצים. הפונקציה ההיפרבולית מאפשרת לרשת ללמוד טווח רחב יותר של ערכים מאשר פונקציית הסיגמואיד.
3. ReLU - ReLU היא אחת מפונקציות ההפעלה הפופולריות ביותר. הפונקציה מחליפה את כל הערכים השליליים בערך אפס, ומאפשרת רק לערכים חיוביים לעבור. הפונקציה הוכחה כיעילה ואפקטיבית מבחינה חישובית. תכונה חיובית של פונקציית אקטיבציה זו היא העובדה שהיא עוזרת להתמודד עם בעיות בגידול הגרדיאנט ולהמנע מתופעות גרדיאנט מתפוצץ.



איור 5 - אילוסטרציה של פונקציית אקטיבציה מסוג סיגמואיד [16].

3.1.3 אימות הרשת

לאחר או במקביל לתהליך האימון של הרשת, מתבצע תהליך **האימות** המתבסס על קבוצת האימות (קבוצת הוולדציה). קבוצה זו היא קבוצת דוגמאות קטנה אשר לרוב נלקחת כחלק מסוים מקבוצת האימון. המודל איננו מתאמן על קבוצה זו ואינו לומד את הדפוסים הייחודיים לה.

תהליך האימות יכול לקרות במקביל לתהליך האימון או מיד לאחריו. בתהליך זה, מוזנות בזו אחר זו דוגמאות מקבוצת האימות לרשת, והמידע מפעפע קדימה עד לשכבת הפלט. לאחר מכן מחושב המרחק בין הפלט של הרשת לתיוג האמיתי של כלל הדוגמאות ובכך נמדדת רמת הדיוק של הרשת או ערכי פונקציית ההפסד (ניתן לחשב מדדי הערכה שונים נוספים במידת הצורך). באמצעות תהליך אימות הרשת תו"כ תהליך האימון ניתן למנוע תהליך של התאמת יתר ולעצור את האימון בנקודות זמן מתאימות. תהליך האימות הוא כלי משמעותי בניסיון למנוע התאמת יתר של רשת הניורונים. הפרדת קבוצת האימות מקבוצת האימון לא תמיד אפשרית. לעיתים מרחב המדגם קטן ומכיוון שבתהליך האימון אין שימוש בקבוצת האימות לצורך אימון המודל, הפרדת תת קבוצה זו יכולה להקטין את סט הנתונים המיועד לאימון הרשת.

אי לכך לעיתים נעשה שימוש בטכניקת אימות המכונה k -fold cross validation. בתהליך זה, קבוצת האימון מחולקת ל- k קבוצות, כאשר בכל איטרציה קבוצה אחת מתוכן נבחרת כקבוצת האימות בעוד שיתר $k-1$ הקבוצות מהוות קבוצת האימון. התהליך מתבצע ב- k איטרציות, ובסופו הדיוק של הרשת מוערך כממוצע מדדי הדיוק שחושבו בכל האיטרציות.

3.1.4 בחינת הרשת

תהליך **המבחן** מתבצע לאחר תהליכי האימון והאימות ומטרתו היא להעריך את ביצועי הרשת. בתהליך זה דוגמאות מקבוצת המבחן (Test set), שלא שימשו עד כה לאימון או לאימות הרשת, מוזנות לרשת בזו אחר זו. גם בתהליך זה מחושב המרחק בין הפלט של הרשת לתיוג האמיתי של הדוגמה, ומחושבים מדדים שונים של הרשת, כמו דיוק, ייחודיות ורגישות, אשר מאפשרים לקבוע את ביצועיה. בפרק חמש אסקור את מדדי ההערכה המרכזיים הנמצאים בשימוש למשימות בחינת רשתות נוירונים.

3.1.5 בעיית התאמת יתר (Overfitting)

מטרת תהליך האימון היא כאמור ליצור מודל אשר מצליח לסווג דוגמאות מתוך סט הנתונים בזהה לתיוג האמיתי על בסיס זיהוי הדפוסים הקיימים בכל דוגמה ובמאגר הנתונים הרחב. בתהליך האימון אנו חושפים למודל מרחב דוגמאות אשר אמור לייצג את העולם האמיתי, אך איננו יכולים לחשוף את המודל לכלל הדוגמאות המרכיבות את העולם האמיתי ואנו מספקים עבורו רק דגימה ממנו.

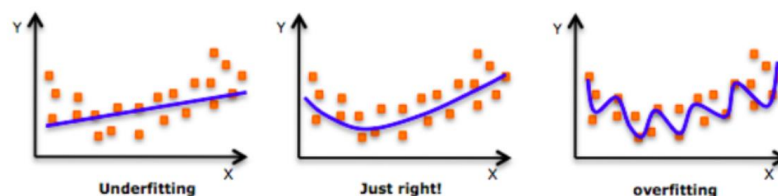
מצב זה מציב אתגר ייחודי שבו אנו רוצים להימנע מתוצאות לא מדויקות של המודל הן מבחינת זיהוי תהליכים לוקה בחסר בעקבות התאמת חסר וזיהוי דפוסים חלקי על ידי המודל

והן מבחינת התאמת יתר של המודל למרחב המדגם. התאמת חסר היא מצב שבו המודל אינו מצליח לזהות דפוסים משמעותיים בקבוצת האימון ולכן לא מצליח ליצור תיוגים נכונים של הדוגמאות המוזנות לו. התאמת יתר היא בעיה מורכבת יותר הנובעת מתהליך האימון של רשתות נוירונים.

מצב של התאמת יתר הוא מצב שבו הרשת מתאימה טוב מדי לדוגמאות בקבוצת האימון, ואינה מצליחה לסווג בדיוק גבוה דוגמאות שאינן משתייכות לקבוצה זו. במילים אחרות, הרשת לומדת דפוסים ספציפיים של קבוצת האימון, אשר אינם רלוונטיים לכלל הדוגמאות בבסיס הנתונים או בעולם האמיתי. במקרה זה, הדיוק של סיווג דוגמאות מקבוצת האימון יהיה גבוה בהרבה מסיווג דוגמאות מקבוצות האימות או המבחן. ככל שהרשת מורכבת יותר וכמות הפרמטרים שהיא צריכה ללמוד גדול יותר, ישנו סיכון להתאמת יתר.

התאמת יתר היא בעיה מוכרת וכדי להתמודד עמה ישנם כלים מסוגים שונים כגון כלי רגולריזציה וכלים העוזרים לעצור את תהליך האימון לפני שהנ"ל מתכנס למצב של התאמת יתר.

איור 9 ממחיש את הבעיות השונות, כאשר בצד ימין שלו מוצגת בעיית Overfitting, כאשר ניתן לראות כי המודל המסווג מתאים לדוגמאות הספציפיות של קבוצת האימון. במרכז האיור ניתן לראות את המודל המסווג שאינו סובל מבעיה זו, שכן הוא מאוזן בסיווג שלו בין הדוגמאות.



איור 6 - השוואה בין אפשרויות להתאמת המודל לבסיס הנתונים *overfitting, underfitting, balanced* [4].

3.1.6 כלים למניעת התאמת יתר והגדלת יכולת ההכללה

בעיית התאמת היתר הינה בעיה מורכבת הפוגעת ביכולות רשת הנוירונים לבצע סיווג מוצלח. לאור הנ"ל פותחו כלים השונים למניעת התאמת יתר בתהליך האימון בניסיון לשפר את יכולת ההכללה והסיווג של רשת הנוירונים [8].

- שימוש בcallback. קולבקים הינן פונקציות אוטומטיות אשר יכולות לעצור את תהליך האימון כאשר מזוהה עליה באיכות הדיוק על קבוצת האימון לצד ירידה או סטגנציה

של רמת הדיוק של קבוצת הוולדציה. נקודת היפרדות זו בערכי רמות הדיוק של קבוצת האימון לעומת קבוצת האימות יכולה להעיד על תחילתה של התאמת יתר בתהליך האימון ולכן עצירת התהליך תביא לערך האופטימלי של פונקציית המשקלים של הרשת.

- דרופאוט dropout. דרופאוט הוא כלי שמטרתו מניעת התאמת יתר על ידי הקטנת התלות של המשקולות בתהליך האימון. בצורה סטטיסטית ובהתאם לערך שנקבע מראש בוחרים מספר משקולות מוגדר מראש שלא ישתתף בתהליך האימון במהלך איטרציה מסוימת במטרה לצמצם את התלות של יתר המשקולות בערכי המשקולות החסרים. בכל איטרציה נבחרת קבוצת משקולות אחרת שבה לא יעשה שימוש לניבוי התייגים באיטרציה נתונה.
- נרמול קלטים – שימוש בשכבות המבצעות נרמול של באץ דוגמאות קלט יכול למנוע חלק מתהליך התאמת היתר שתהליך האימון.

בנוסף לכלים אלו במחקרים רבים מבוצע שימוש באוגמנטציה

אוגמנטציה היא תהליך עיבוד מקדים שמטרתו הגדלת כמות הדוגמאות שבעזרתן ניתן לאמן את רשת הנוירונים. כאשר אנו מאמנים רשת נוירונים, כמות הדוגמאות שהן אנו משתמשים היא קריטית לתהליך האימון. ככול שמספר הדוגמאות גדול ומגוון יותר כך באפשרותנו ללמד את המודל דפוסים מורכבים יותר ולשפר את יכולת ההכללה שלו. מאגר מידע גדול יותר מעט אף תהליך התאמת היתר כיוון שהמודל מנסה למצוא דפוסים בקרב מספר דוגמאות גדול יותר [8].

יחד עם זאת, הוספת דוגמאות למאגר המידע שבו אנו משתמשים הוא לא משימה פשוטה תמיד. במימד הרפואי דאטה מתווייג הוא יקר במיוחד – לרוב אנו נזדקק לחוות דעתו של רופא מומחה לצורך סיווג מידע חדש ולא תמיד נוכל למצוא מאגרי מידע מתאימים לסוגיית המחקר.

בכדי לפתור את הבעיות הנ"ל מבוצע שימוש באוגמנטציה. במקום ליצור מאגר מידע חדש ולעסוק בתיג דוגמאות חדשות, אנו משנים מעט את מאגר המידע שברשותנו באמצעים שונים בכדי לשנות ולהגדיל את מאגר המידע בדרכים מלאכותיות. כלי אוגמנטציה יכולים לבצע מניפולציות רבות על מאגר המידע בכדי ליצור שינויים במאגר ובכך ליצור דוגמאות נוספות לאימון. בין המניפולציות שניתן לבצע על מאגר המידע ניתן למצוא סיבוב הדוגמאות בכיוונים שונים, היפוך אופקי או אנכי, הבהרה או החשכה של גוון הדוגמה, הזזה, מתיחה או כיווץ של הדוגמה וכלים נוספים [8].

3.2 רשתות קונבולוציה

בפרק הקודם סקרתי רשתות נוירונים עמוקות מסוג fully connected (FC) שבהן כל נוירון מחובר לכל הנוירונים בשכבה שלאחריו ומלפניו. לשיטה זו מספר מגרעות. ראשית היא דורשת כוח חישובי גדול. כל קשר בין זוג נוירונים מתורגם למשקולת שערכה מתעדכן בתהליך הלמידה. עדכון המשקולות בתהליך הלימוד דורש חישוב של הנגזרת החלקית של כל משקולת וחישוב גודל התיקון הנדרש עבור כל משקולת בתהליך האופטימיזציה. ככול שמספר המשקולות ברשת גדול יותר כך דרושות מספר רב יותר של פעולות גזירה, חישוב ועדכון ערך המשקולות. מלבד תהליך עדכון המשקולות, פעולת הפעפוע לפנים דורשת עוצמה חישובית גדולה יותר כיוון שכמות החישובים בשכבות השונות גדלה בהתאם לכמות המשקולות המחברות את הנוירונים [14].

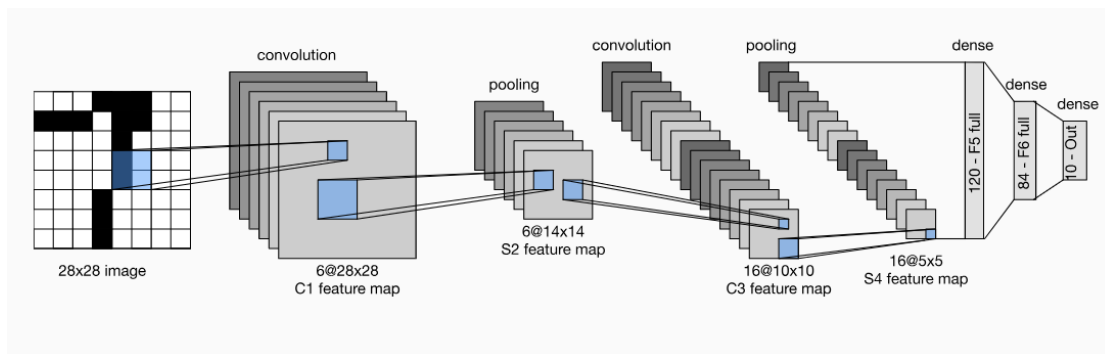
מלבד סוגיית העוצמה החישובית לעיתים אין צורך בכמות הקשרים הגדולה בין כלל הנוירונים בין כלל שכבות הרשת. לא תמיד יש קשר בין כלל איברי הקלט. כאשר עוסקים בתמונות חזותיות לעיתים קרובות אין קשר בין פיקסלים הנמצאים במקומות רחוקים אלו מאלו ולכן אין חשיבות בחיבור כלל שכבות הקלט זו לזו. ניתן להניח שבתמונות ובאמצעים חזותיים ישנם קשרים "מקומיים" בין פיקסלים צמודים אך לרוב אין קשר הדוק בין צד מסוים של התמונה לצידה השני. אי לכך אין טעם לבזבז עוצמה חישובית בניסיון למצוא קשרים בין חלקים מופרדים בתמונות ויש להתמקד בקשרים הנוצרים בסמיכות הגאוגרפית בין פיקסלים סמוכים ואזורים קרובים בתמונות הקלט. כדי להימנע מהבעיות הנ"ל ניתן להשתמש ברשתות קונבולוציה [14].

רשתות עצביות קונבולוציוניות (CNNs) הן סוג מסוים של רשתות נוירונים המשמשות בעיקר לטובת עיבוד וניתוח נתונים חזותיים. רשתות מיוחדות אלו הוכיחו אחוזי דיוק גבוהים במשימות כגון ניתוח ועיבוד תמונה, זיהוי אובייקטים וסיווג תמונות [22].

רשת קונבולוציה מקבלת כקלט מטריצה אשר לרוב מייצגת תמונה, וכפלט הרשת יכולה לזהות ולסווג אובייקטים בתוכה או לפענח ולעבד את מרכיביה של התמונה. הארכיטקטורה של רשת קונבולוציה נבנתה בהשראת המבנה של הקורטקס הוויזואלי, האזור המרכזי במוח בו מעובד מידע חזותי הנקלט במערכת הראייה. כל נוירון מגיב לגירוי הוויזואלי רק באזור מוגבל בלבד, כך שאוסף האזורים הנקלטים על ידי כל הנוירונים מכסים את כלל שדה הראיה.

באיור 10 ניתן לראות מבנה של רשת CNN טיפוסית. הרשת מקבלת כקלט תמונה המיוצגת כמטריצת פיקסלים בפורמט רב מימדי כגון RGB או בפורמט חד מימדי כגון גווני אפור או שחור-לבן. הרשת מפעילה מספר שכבות מסוגים שונים עליהן נעמוד בהמשך. ניתן לחלק את הרשת לשני חלקים מרכזיים בצורה סכמטית. בחלק הראשון של הרשת מתבצעת למידת

המאפיינים (Feature Learning) של תמונת הקלט. בחלק זה מופעלות שכבות קונבולוציה ואיגום (Pooling), שמטרתן לחלץ את המאפיינים החשובים מתמונת הקלט תוך הקטנת ממדי התמונה. בתום חלק זה, מפת המאפיינים שמתקבלת מוזנת לרשת נוירונים מלאכותית נוספת אשר מבצעת את משימת הסיווג של התמונה לאחת ממחלקות הסיווג עבור הבעיה הנתונה. באופן עקרוני, מטרת פעולות הקונבולוציה והאיגום היא להקטין את ממדי התמונה תוך שמירה על המאפיינים החשובים ביותר שלה. בנוסף פעולות אלו נועדו על מנת לצמצם את הכוח החישובי הנדרש לרשת הנוירונים לבצע את פעולת הסיווג [8].



איור 7 - מבנה רשת קונבולוציה לדוגמה [8].

יתרונה של רשת קונבולוציה אם כן הוא כפול – גם חסכון במשאבים וגם התייעלות ושיפור ביצועים בהתאם למישור הבעיה החזותית. אדון עתה בשכבות האופייניות לרשתות קונבולוציה.

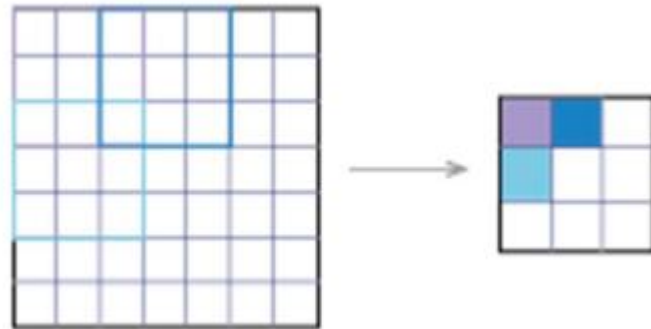
3.2.1 שכבות אופייניות לרשתות קונבולוציה

רכיב הבסיס של רשת קונבולוציה היא **שכבת הקונבולוציה**. השכבה מבצעת קונבולוציה לינארית על פני הקלט בכדי לקבל ייצוג פשוט יותר שלו. שכבת הקונבולוציה מבצעת פעולת קרוס-קורלציה בין וקטור המשקולות ובין הקלט (הקלט ברשת יהיה וקטור הכניסה לרשת או וקטור היוצא משכבה חבויה כלשהיא בתוך הרשת). וקטור המשקלים נקרא גרעין הקונבולוציה או מסנן. פעולת הקרוס קורלציה מבוטאת כך:

$$y[n] = \sum_{m=1}^{k-1} x[n-m]w[m]$$

משוואה 2 - משוואת קרוס קורלציה [14].

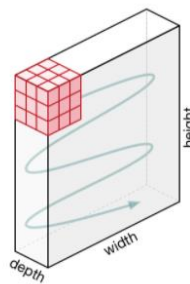
כאשר x הוא ווקטור הכניסה $w[i]$ הוא ווקטור המשקלים. גודל הגרעין יהיה קטן יחסית לתמונת הקלט. פעולת הקונבולוציה מתבצעת כך שהגרעין "מונח" על גבי חלק מהתמונה ומחושבת המכפלה הסקלרית של הגרעין עם החלק בתמונה שהוא מכסה.



איור 8 - פעולת הקונבולוציה [14].

באיור 11 ניתן לראות דוגמה לפעולת הקונבולוציה. מצד שמאל מתוארת תמונת הקלט ומצד ימין תוצאות פעולת הקונבולוציה. ניתן לראות כי עבור כל מיקום של הגרעין לאורך תמונת הקלט (מיקום הגרעין על פני תמונת הקלט מודגש בצבעים שונים), נוצר ערך בודד בתמונת הפלט. התקדמות הגרעין על גבי התמונה מתחילה מהפינה השמאלית העליונה לכיוון הפינה הימנית התחתונה, כאשר קצב התקדמות הגרעין (Stride) קובע את כמות הפיקסלים שיזוז הגרעין בין צעד לצעד.

איור 12 מתאר באופן סכמתי את אופן התקדמות הגרעין:



איור 9 - התקדמות הגרעין על פני תמונת הקלט בפעולת הקונבולוציה; הגרעין מתחיל מהפינה השמאלית העליונה ומתקדם לכיוון הפינה הימנית התחתונה של התמונה [21].

ברשתות קונבולוציה ישנן לרוב מספר שכבות קונבולוציה. כל שכבה מקבלת כקלט תמונה, ומבצעת עליה קונבולוציה באופן שתואר לעיל. התמונה המתקבלת מכונה **מפת המאפיינים**. בדומה לרשתות נוירונים מלאכותיות, גם כאן מופעלת על כל אחד מהערכים של מפת המאפיינים פונקציית אקטיבציה ורק אז היא מועברת לשכבות הבאות ברשת.

מלבד הקטנת כמות המשקלים הנלמדת, שכבת הקונבולוציה עוזרת לזיהוי דפוסים המבוססים על סמיכות גאוגרפית במרחב התמונה. יכולת זו נובעת מפעולת הקונבולוציה אשר בודקת חפיפה בין הגרעין לבין חלקים משכבת הקלט. ישנם סוגים שונים של גרעינים היכולים לבצע פעולות שונות כגון החלקה או טשטוש של תמונת הקלט. בכל שכבת קונבולוציה נמצאים מספר רב של גרעינים אשר כל אחד מהם לומד בתורו מאפיינים שונים של תמונת הקלט. ככול שהרשת בעלת שכבות חבויות רבות כך הרשת תוכל ללמוד דפוסים מובחנים יותר של תמונת הקלט [14].

כאמור, הקלט של רשת הקונבולוציה יכול להיות רב מימדי (לדוגמה קלט צבעוני בפורמט RGB), במקרה זה פעולת הקונבולוציה יכולה להתבצע על כל ערוצי הקלט ולספק ערוץ פלט חד מימדי, או לבצע פעולה על כל ערוץ בנפרד ולייצר פלט רב מימדי.

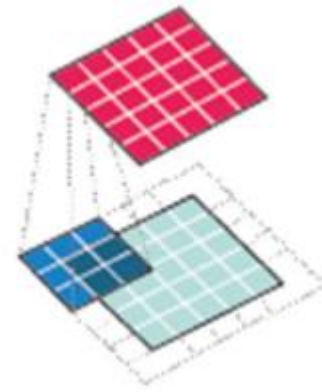
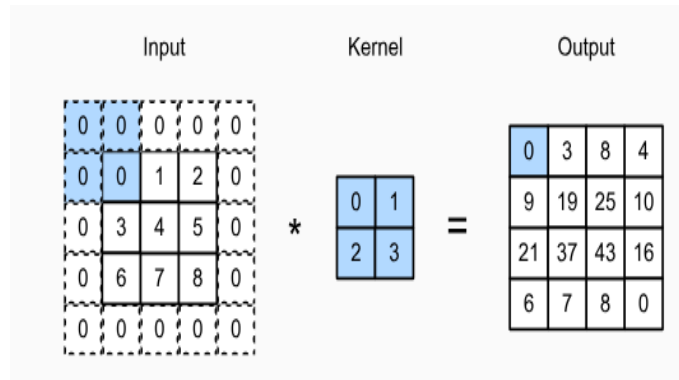
בנוסף בכל שכבת קונבולוציה ניתן ליצור כמות מוגדרת של גרעינים כך שכל גרעין "ינוע" על גבי תמונת הקלט בצורה עצמאית ובלתי תלויה בשאר הגרעינים. בצורה זו ניתן ליצור מספר רב של שכבות פלט, כאשר כל שכבה היא תוצאה של פעולת גרעין נתון על תמונת הקלט.

ריפוד (Padding)

פעולת הקונבולוציה המבוצעת על ידי הגרעין הינה פעולה מרחבית, כלומר הגרעין פועל על קבוצת איברים בכל פעולה. גודל סטנדרטי של גרעינים נע בין מטריצות בגודל 3X3 ונע גם עד גדלים של 7X7 ויותר.

נשים לב כי פעולת הקונבולוציה לא מוגדרת על איברי הקצוות של תמונת הקלט כיוון שתנועת הגרעין לאורך תמונת הקלט מוגבלת יחסית למיקום מרכז הגרעין. ולכן תמונת הפלט מהפעלת הגרעין היא תמונה בעלת מימדים קטנים יותר יחסית לתמונת הקלט. במידה ואנו רוצים לבצע את פעולת הקונבולוציה גם על איברי הקצוות נדרש "לרפד" את קצוות התמונה בערכים מלאכותיים. קיימות שיטות שונות לביצוע ריפוד ולקביעת ערכי המספרים שבהם יתבצע שימוש למילוי הערכים המרופדים. בחלק מהמקרים ממלאים את המקומות הנדרשים לריפוד בערכי אפס ובמקרים אחרים משכפלים את ערכי הקצוות לצורך הריפוד.

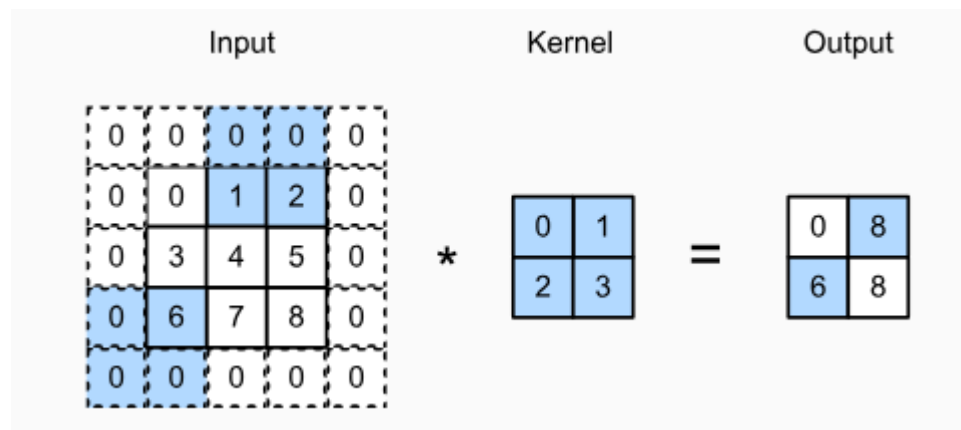
גודל שכבת הריפוד משתנה אף היא בהתאם לצורך ויכולה להכיל יותר משורה או טור ריפוד בודד.



איור 10 - פעולת הריפוד. משמאל פעולת הקונבולוציה ללא ריפוד. ניתן לראות שגודל תמונת הפלט קטן יותר בהתאם לגודל הגרעין. מימין הקלט המקורי מקבל ערכי ריפוד (מופיעים בקווים דקוקים דהויים) בכדי שפעולת הקונבולוציה תתקיים גם על דפנות הקלט [14].

גודל הצעד (stride)

על מנת להקטין את העוצמה החישובית הדרושה ברשתות קונבולוציה ניתן להפעיל את פעולת הקונבולוציה בצורה דלילה יותר מאשר על כל רכיב בתמונת הקלט. הנחה זו נובעת מהתפיסה כי לרוב נשמר הקשר המרחבי באזורים קרובים בתמונה ולכן אין צורך להשתמש בגרעין על כלל האזורים בתמונת הקלט וניתן לבצע דילוגים במרחבים מצומצמים. מרחק הפסיעה של הפעלת הגרעין על אזורים בתמונת הקלט נקרא Stride. ככול שגודל הצעד גדול יותר כך גדל המרווח במיקום הפעלת הגרעין על גבי תמונת הקלט.



איור 11 - שימוש בגודל צעד משתנה במישור האופקי ובמישור האנכי [8].

פעולת האיגום (Pooling)

הפעולה מתבצעת כאשר גרעין מועבר על גבי מפת המאפיינים, ובכל שלב מתבצע חישוב באזור עליו מונח הגרעין, אשר נכנס לפלט של שכבת ה-Pooling. באיור מטה ניתן לראות כי

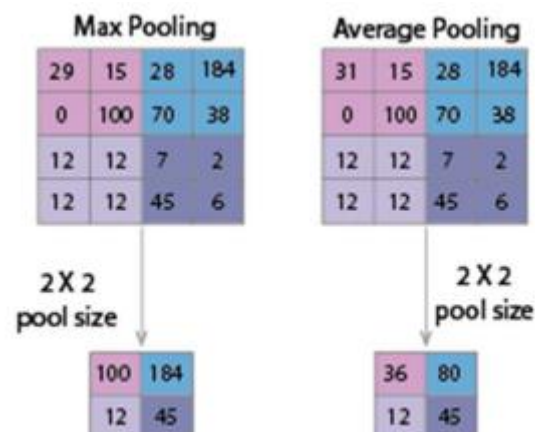
בחירת ערך המקסימום מתוך החלק הכהה במטריצה הימנית מניב את התוצאה 2, אשר מופיעה בחלק הכהה של המטריצה השמאלית. כך מועבר הגרעין על פני כל המטריצה הימנית, וערך המקסימום בכל שלב נלקח למטריצה השמאלית.

3.0	3.0	3.0
3.0	3.0	3.0
3.0	2.0	3.0

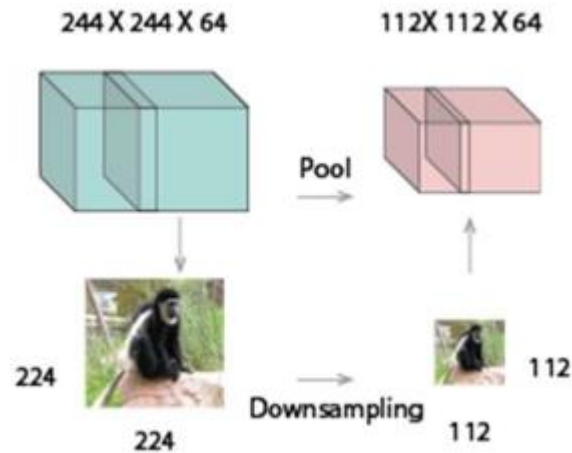
3	3	2	1	0
0	0	1	3	1
3	1	2	2	3
2	0	0	2	2
2	0	0	0	1

איור 12 - פעולת איגום (Pooling); הפעולה מתבצעת על תמונת הקלט בגודל 5×5 (מימין) עם גרעין בגודל 3×3 . המטריצה משמאל היא מטריצת הפלט של הפעולה, והתא המסומן באפור בה הוא התוצאה של בחירת ערך המקסימום מאותו אזור המסומן באפור בתמונת הקלט

ישנן שתי שיטות עיקריות לביצוע Pooling, אשר מתוארות באיור 16. האחת מכונה Max Pooling, ובה נלקח ערך המקסימום מכל חלק בגודל מוגדר של מפת המאפיינים (באיור תת מרחב בגודל 2×2), והשנייה מכונה Average Pooling, ובה נלקח הערך הממוצע מכל חלק בגודל מוגדר של מפת המאפיינים:



איור 13 - פעולת האיגום. מימין איגום בשיטת AVERAGE ומשמאל איגום בצורת MAX [14].



איור 14 - הפעלת פעולת האיגום על מישור התמונה [14].

Fully connected Layer

שכבות מלאות או שכבות מחוברות לחלוטין הן חלק מרשת הקונבולוציה השלמה ומשמשות בעיקר בכדי לטייב את איכות יכולת הסיווג של הרשת. שכבות אלו מחברות כל נוירון לכל נוירון אחר בשכבות סמוכות, ומאפשרות לרשת לעבד את התכונות שחולצו מרשת הקונבולוציה ולהפיק את פלט הסיווג של הרשת.

3.2.2 אימון אימות ומבחן הרשת

ככלל תהליכי האימון, אימות ובחינת רשת קונבולוציה דומים לאימון רשת FC.

הארכיטקטורה של הרשת נקבעת מראש על ידי גדלי הגרעינים המשמשים בשכבות הקונבולוציה והאיגום, על ידי פונקציות האקטיבציה ועל ידי רשת הנוירונים המלאכותית. במהלך אימון הרשת המשקלים הנדרשים ללמידה הינם הערכים של גרעיני הקונבולוציה והמשקלים של רשת הנוירונים המלאכותית. גודל הפסיעה, גדלי הגרעינים, סוג האיגום המבוצע וסוג הריפוד הם גדלים ומשתנים קבועים שאינם משתנים לאורך הרשת.

גם ברשת מסוג זה, נדרש לקבוע אלגוריתם אקטיבציה כגון Gradient Descent, כאשר בכל שלב מוזנת תמונה מבסיס הנתונים לרשת, המידע מפעפע קדימה (Forward Propagation) ולאחר מכן מפעפע אחורה (Back Propagation) על מנת לחשב את גרדיינט פונקציית המחר, כפי שתואר בסעיף 3.1.3 בעבודה זו.

תהליך האימות מבוצע בצורה דומה אף הוא. בתהליך זה מבוצע שימוש בדוגמאות שאותן לא למד המודל בכדי לבצע תהליך אימות. גם ברשת קונבולוציה ניתן לעצור את תהליך האימות משמזוהה התאמת יתר בתהליך האימות. כך הדבר נכון גם עבור מבחן הרשת.

3.2.3 רשתות קונבולוציה הנמצאות בשימוש נרחב

ישנם סוגים רבים של רשתות קונבולוציה מוכרות שפורסמו על ידי הקהילה המחקרית. רשתות אלו מפגינות יכולות מעולות באדפטציה בין משימות שונות בעולמות תוכן שונים. בין הרשתות המוכרות והנפוצות בעולם ניתן למצוא:

1. VGG – רשת שפותחה בשנת 2014 על ידי מדענים מאוניברסיטת אוקספורד והייתה הרשת הראשונה שהשתמשה בפילטרים שגודלם 3×3 בכל שכבת קונבולוציה. באמצעות ארכיטקטורת VGG הוכיחו שהגדלת מספר השכבות ברשת מובילה לשיפור בתוצאותיה. החיסרון העיקרי ברשת VGG הוא מספר הפרמטרים העצום הנובע מהארכיטקטורה עצמה ואשר גרם לצריכת זיכרון גדולה וירידה בביצועים.
2. DenseNet – ארכיטקטורת רשת זו שואבת השראה מארכיטקטורת ה-ResNet. בדומה ל-ResNet, גם רשת זו בנויה מבלוקים אשר נקראים Dense Blocks. בארכיטקטורה זו כל קבוצת שכבות מתחברת לכל השכבות האחרות. ביצועי הרשת דומים מאוד לביצועי ResNet, אך ב-DenseNet מספר הפרמטרים הוא כמחצית ממספרם ב-ResNet, ולכן הוא מפחית את צריכת הזיכרון ומסייע בשיפור בביצועים.
3. ResNet היא שם של מגוון רשתות קונבולוציוניות בגדלים שונים כגון 50RESNET או 16 RESNET. לרשתות אלו ישנן מספר תכונות עיקריות אשר משפרות את ביצועיה מעבר לפעולות הקונבולוציה הבסיסיות [20]:

1. חיבורים שיוריים – חיבורים אלו הן אחת מהתכונות הבסיסיות של הרשת. חיבורים אלו מאפשרים לרשת "לדלג" על שכבה אחת או יותר במהלך האימות, וליצור למעשה קיצורי דרך או "דילוג על חיבורים", בתהליך הלמידה. בצורה זו מתמודדת הרשת עם בעיית השיפוע הנעלם, ומאפשר לאמן רשתות עמוקות מאוד ללא ירידה בביצועים. זמן הריצה מושפע בשני אופנים. ראשית זמן ההתכנסות של המודל לנקודת אופטימום היא מהירה יותר. עובדה זו נובעת מכך שרשת הנירונים לומדת בצורה אפקטיבית יותר. זמן הריצה של כל אצווה הוא מעט איטי יותר, כיוון שכמות הפרמטרים שהרשת נדרשת ללמוד הוא גדול יותר. יחד עם זאת, וכיוון שההשפעה של זמן התכנסות מהיר יותר עולה על ההשפעה

של זמן האימון הגדול יותר בכל אצווה, בסה"כ זמן לימוד הרשת בכללותו מהיר יותר.

2. עומק רשת - כפי שהשם מרמז, רשתות resnet מורכבת ממספר שכבות משתנה

בהתאם לשמן. העומק של הרשתות הוא גורם מפתח בהצלחתן המאפשר לרשת

ללמוד ייצוגים היררכיים מורכבים של תכונות בתמונות קלט. הארכיטקטורה

העמוקה יותר מאפשרת ל-ResNet-101 ללכוד דפוסים מופשטים ומורכבים יותר,

חיוניים למשימות כמו סיווג תמונה וזיהוי אובייקטים.

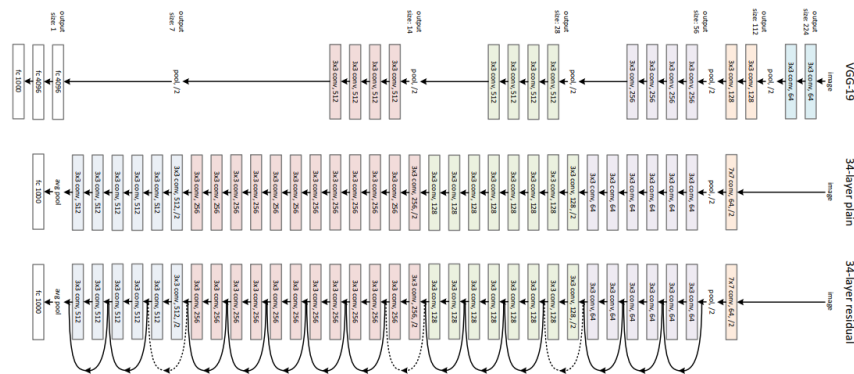
3. איגוד ממוצע "עולמי" - במקום שכבות מסורתיות מחוברות במלואן, resnet

משתמשת בדרך כלל באיגום ממוצע גלובלי בשכבות האחרונות. טכניקה זו

מבצעת ממוצע של הערכים של כל מפת תכונות, ומספקת ייצוג קומפקטי ויעיל

של התכונות הנלמדות. איגום ממוצע עולמי מפחית את הסיכון להתאמת יתר

ותורם ליכולת ההכללה.



איור 15 - דוגמה מתוך רשת RESNET 101 [20].

3.3 רשתות חוזרות (RNN)

רשתות נשנות (RNNs) הן חידוש מרכזי בתחום הלמידה העמוקה ונועדו לשימוש עבור עיבוד נתונים עוקבים או נתונים סדרתיים. שלא כמו רשתות עצביות מסורתיות ובשונה מרשתות קונבולוציה שיתרון נמצא בכך שהן מסוגלות לעבד ולנתח קלט חזותי ותמונות RNNs יש יכולת לשמור מידע מפרקי זמן קודמים ממרחב הקלט ולהשתמש במידע שנשמר בכדי לשפר ולטייב את העיבוד והפקת המידע מסדרת הנתונים. יכולת זו הופכת את הרשתות הנשנות למתאימות במיוחד למשימות כמו עיבוד שפה טבעית וניתוח סדרות נתונים [3].

אבן הבניין הבסיסית של רשת נשנית הוא תא הרשת העצבי הנשנה. ייחודו של תא זה הוא היכולת לעבד ולשמור מידע הנצבר ונוצר משלבים קודמים בסדרת הקלט יחסית לחלק הנוכחי הנמצא בשלב העיבוד בתוך התא מבין רצף הנתונים המוזן לרשת. היכולת לשמור מידע רלוונטי מרצף נתוני העבר מאפשר לתא הרשת ללמוד הקשרים ותלותיות בתוך רצף הנתונים. ישנם סוגים שונים של תאי זיכרון המבטאים יכולות ועוצמות חישוב משתנות. רשתות נוירונים נשנות נעות על גבי רצפים וסדרות (או על פני שלבי זמן) כך שאותם פרמטרים ומשקולות מיושמים מספר פעמים ובהתאם לאורך הקלט [8].

רשת נשנית בנויה לרוב משילוב של תאי זיכרון לצד רשתות מחוברות מסוג FC. גם כאן בזהה לרשתות הקונבולוציה, תפקידם של תאי הזיכרון הוא לשמש כתאים המעבדים ומנתחים את המאפיינים של קלט הנתונים מבוסס הרצף, ופלט תאי הזיכרון מוזן לרשת FC לטובת משימת סיווג או חיזוי נתוני הפלט של המודל. בשונה מרשת קונבולוציה, תאי הזיכרון מעבדים מידע המבוסס על רצף המשתנה לאורך זמן ולא על בסיס מיקום המידע על גבי מישור תמונה או פריים.

ככלל ניתן לתאר את מרכיבי המפתח של תא RNN באופן הבא [3]:

1. שכבת קלט - שכבת הקלט של תא RNN מעבדת את הקלט בשלב הזמן הנוכחי.
2. מצב נסתר - המצב הנסתר מכיל בתוכו את המידע שנלמד משלבי הזמן הקודמים של הקלט. המצב הנסתר משמש כרכיב הזיכרון של התא ושומר מידע נחוץ לעיבוד הקלט ומאפשר לרשת ללמוד תלויות והקשרים הנמצאים ברצף הנתונים.
3. מטריצות משקל - מטריצות משקל קובעות את השפעת הקלט והמצב הנסתר על הפלט הנוכחי. משקלים אלו נלמדים במהלך תהליך האימון, ומאפשרים לרשת להסתגל לדפוסים הספציפיים בתוך הנתונים.
4. פונקציית אקטיבציה - לעתים קרובות פונקציה סיגמואידית המאפשרת ללכוד יחסים מורכבים בתוך הנתונים.

5. שכבת פלט - הפלט הוא שילוב של הקלט הנוכחי והמידע המאוחסן במצב נסתר, אשר מייצר את החיזוי או הייצוג לשלב הזמן הנוכחי.

תאי RNN מאפשרים לתת חשיבות לסדר של האיברים המוכנסים אליהם. ניתן להסתכל על המשקולות הפנימיות של תאי RNN כאל רכיב זיכרון פנימי. כאשר כל מרכיב שנכנס אל תא RNN משוקלל בפונקציה פנימית ביחס לרכיב פנימי המשתנה בהתאם לערכי רכיבי העבר.

כאשר דנו ברשתות נוירונים מחוברות במלואן או ברשתות קונבולוציה התייחסנו לכל פיצ'ר $X \in \mathbb{R}^2$ כאל ערך יחיד. בבואנו לדון ברשתות נשנות עלינו להתייחס לכל אחד מהפיצ'רים בדוגמה הקלט כאל סדרה של ערכים $x_1, x_2, x_3, \dots, x_T$ כאשר $T \in \mathbb{Z} +$ הוא כתלות במיקום שלו בסדרה.

כאשר נכנס אינדקס x_t מסוים לתא הזיכרון הוא מוכפל במשקולת w_{xh} ונאגר בתוך אינדקס הזיכרון h_t . האינדקס h_t הוא פונקציה של x_t ו- h_{t-1} :

$$h_t = f(h_{t-1}, x_t)$$

משוואה 3 [14].

כך המידע לגבי הקלט בנקודת הזמן t המיוצגת על ידי האינדקס h_t מושפעת גם מהמידע שנצבר בתא הזיכרון עד כה המיוצג על ידי האינדקס h_{t-1} וגם על ידי ערך הקלט הנוכחי x_t .

מלבד המשקלים הפועלים על ווקטור הכניסה, ישנם גם משקלים הפועלים על הרכיבים הפנימיים (משקולות הזיכרון), ומשקלים הפועלים גם על המוצא של האינדקס w_{hy} .

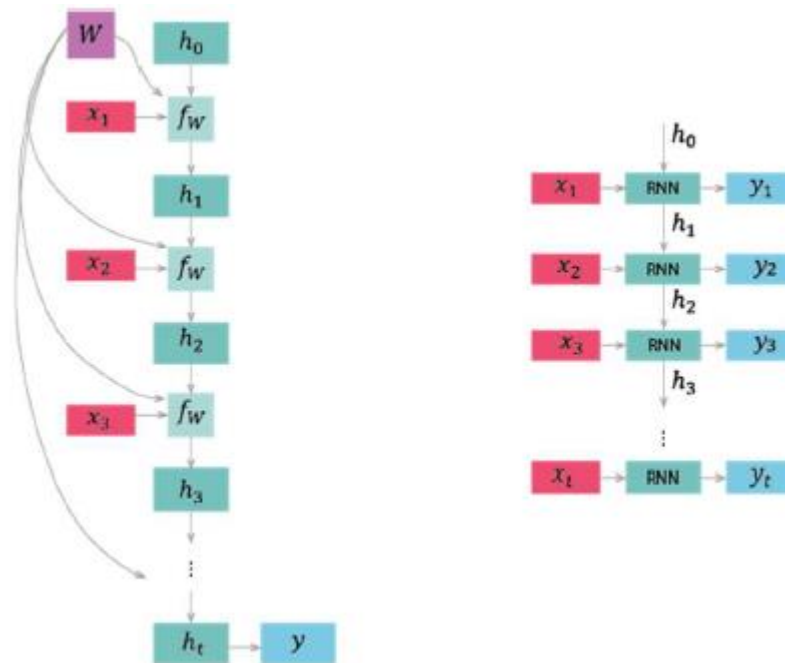
המשקלים z הים ומתעדכנים ביחד. בנוסף ניתן להשתמש בפונקציית אקטיבציה f מסוגים שונים (סיגמואיד, Relu וכו').

באופן פורמלי, ניתן להביע את התהליך כך:

$$h_t = f_w(w_{hh}h_{t-1} + w_{xh}x_t)$$

משוואה 4 [14].

ניתן לתאר את התהליך גם באופן סכמתי כפי שמופיע באיור מטה:



איור 16 - ארכיטקטורת RNN בסיסית [14].

הפלט של תאי RNN הוא ווקטור הנקרא ווקטור ההקשר. ווקטור זה (קרי המצב הנסתר של התא לאחר הכנסת החלק האחרון של הווקטור x אל הרשת) מייצג את המידע שהרשת "זכרה" מקלט הרצף. ווקטור ההקשר אמור להכיל ייצוג של כל התלויות המשמעותיות המוכלות בתוך הנתונים הגולמיים והוא תוצאה של העיבוד שבוצע בתוך תאי RNN. כלומר הווקטור מכיל בתוכו את ביטויי התלויות המשמעותיות ביותר בקלט. ווקטור ההקשר מועבר לרוב להמשך עיבוד ברשת הנירונים אם לצורך סיווג או חיזוי [3].

3.3.1 תהליך האימון של הרשת

הלמידה של הרשת נעשית בצורה דומה לרשתות ניורונים עמוקות עם מספר סייגים.

נגדיר את פונקציית המחיר של רשת RNN:

$$L(\theta) = \frac{1}{n} \sum_i L(y_i^{\wedge}, y_i, \theta)$$

משוואה 5 [14].

נוכל להשתמש באלגוריתם GD רגיל אך לא נוכל לבצע את תהליך הפעפוע לאחר הרגיל כיוון שכל משקל מופיע ברשת מספר פעמים. למשל w_{hx} פועל על כל האיברים ברצף הקלט, ובנוסף w_{hh} פועל על כל רכיבי הזיכרון. כדי לבצע עדכון של המשקלים נדרש לבצע תיקון של תהליך הפעפוע לאחר. אנו נסתכל על הרשת הנפרשת כרשת אחת גדולה, ונחשב את

הגרדיאנט עבור כל משקל ומשקל לאחר מכן נסכום את הגרדיאנטים ונבצע שקלול או מיצוע [14].

אם המידע שנכנס לרשת הוא בגודל n קרי יש n דגימות זמן, אז ישנם גם n רכיבי זיכרון ו- n משקולות.

ולכן הגרדיאנט המשוקלל יהיה:

$$\frac{\partial L}{\partial W_{hh}} = \sum_{n=1} \frac{\partial L}{\partial W_{hh}(t)} \text{ or } \frac{\partial L}{\partial W_{hh}} = \frac{1}{n-1} \sum_{n=1} \frac{\partial L}{\partial W_{hh}(t)}$$

משוואה 6 [14].

במילים אחרות, כדי להתמודד עם האתגר שמציב לנו תא רשת נשנית שעל פיו אותן משקולות נמצאות בשימוש רב פעמי כתלות במספר פעימות הזמן של רצף נתוני הקלט, נדרש לתקן ולהתאים את תהליך הפעפוע לאחר. אי לכך אנחנו מחשבים את הגרדיאנט עבור כל פעימת זמן בנפרד ולאחר מכן מבצעים סכימה או מיצוע של כלל התיקונים הנדרשים לכל משקולת לפני תיקונה [14].

3.3.2 אימות ומבחן רשת נשנה

אימות הרשת ובחינתה מבוצע בצורה דומה לאימון ומבחן רשת קונבולוציה או רשת FC. כיוון שהמידע המוזן לרשת נשנית הוא רצף ארוך של נתונים המשתנה לאורך זמן, פיצול הדאטה לקבוצת מבחן, וולדציה וקבוצת אימון צריך להיות מותאם לרצף הנתונים המזין את רשת הנוירונים בתהליכים השונים.

אי לכך, בכדי לפצל את הדאטה לקבוצות השונות, ניתן לקטוע את רצף הנתונים הרציף ולחלקו לתתי קבוצות באורכים המאפשרים ביצוע אימות ובחינה של רשת הנוירונים. פיצול הדאטה לתתי מקטעים יכול להתבצע בצורה ידנית, קרי על פי אורך הנקבע על ידי החוקר בצורה שרירותית, או באמצעות או באמצעות להיקבע מראש על ידי החוקר, לדוגמה ניתן לקבוע כי מחקר יאמן רשת נוירונים שהקלט שלה הוא באורך קבוע כגון מספר מילים קבוע במשפט, או מספר פעימות זמן קבועות לטובת חיזוי ערך של מניה.

בזמן פיצול הדאטה לקבוצת האימון, האימות והמבחן נוכל לחלק את הפיצ'ר X לאורכים שונים לפי מיקום T לאורך הסדרה. פיצול הקלט לתתי סדרות צריך להיות מותאם לתהליך האימון ולאפשר למודל ללמוד את ההקשרים והתלויות הקיימים בקלט.

3.3.3 ארכיטקטורת הרשת הנשנת

ישנם מודלים שונים של רשתות נשנות המתאימים לבעיות שונות:

One to many – במודל זה ישנו קלט אך ישנה סדרה של פלטים. למשל כאשר אנו מכניסים לרשת תמונה או מילה ואנו רוצים שהרשת תתאר את התמונה באמצעות רשת של מילים.

Many to one – במודל זה קיים קלט סדרתי רב נתונים והפלט של המודל הוא נתון בודד. למשל כאשר הרשת מקבלת משפט המכיל מספר מילים והפלט של הרשת הוא הסיגמנט של המשפט.

Many to many – במודל זה עבור כל סדרת קלט יש סדרת פלט, למשל תרגום משפה אחת לשפה אחרת.

3.3.4 בעיית הגרדיאנט הנעלם והמתפוצץ

בעוד שרשתות RNNs מסורתיות הן רבות עוצמה, ישנם אתגרים רבים בשימוש בהן כאשר מתמודדים עם רצפים ארוכים של מידע. סוגי הבעיות הנפוצות הנוגעות לרשתות RNN הן בעיית השיפוע הנעלם ובעיית השיפוע המתפוצץ. למרות שבעיות מסוג זה הן נפוצות במגוון רחב של רשתות נוירונים, בעיה זו מתעצמת במיוחד בארכיטקטורת רשתות נוירונים נשנות מסורתיות. בעיית השיפוע הנעלם היא בעיה שבה החשיבות של פרט מידע בשלב מוקדם דועכת בשלבים מתקדמים. מקור הבעיה הוא בכך שבכדי לחשב את הגרדיאנט יש לחשב מכפלה של מספרים רבים שערכם קטן מ-1 ומספר המכפילים גדל ככול שאורך הרצף גדל. תופעה זו מקנה לרשת את היכולת "להעדיף" אירועים חדשים ברצף הקלט על פני אירועים ישנים.

בעיית השיפוע המתפוצץ היא הבעיה ההופכית שבה מתבצעת מכפלה של מספרים הגדולים ממש מהספרה 1 – וערכי הגרדיאנט עולים במהירות. כתוצאה מכך ערכי המשקולות מתעדכנים בקצבים גדולים ואין יכולת למצוא את נקודת האופטימום ואף לגרום לקריסת המודל.

מגבלה זו סללה את הדרך לארכיטקטורות מתוחכמות יותר, אחת מהן היא רשת הזיכרון לטווח קצר ארוך (LSTM) [14].

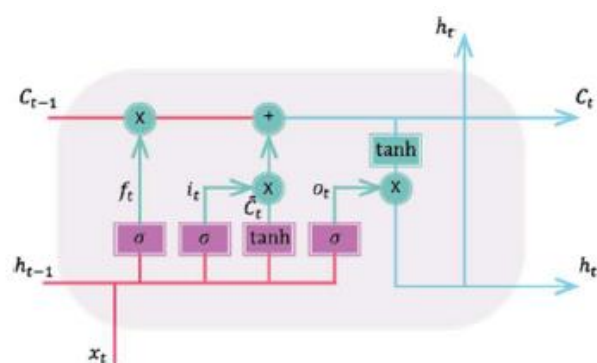
3.3.5 תא LSTM

כדי להתגבר על בעיית דעיכת הגרדיאנט המונע מהרשת ללמוד דפוסים ארוכי טווח, ניתן להוסיף אלמנטים לתאי הזיכרון כך שיוכלו לא רק ללמוד דפוסים מהעבר אלא גם להיות בשליטה על איך וכמה להשתמש בדפוסים הנ"ל. תא LSTM מציע שני שינויים עיקריים יחסית לתא RNN בכדי לנסות ולהתמודד עם הבעיה.

ראשית לתא יש כניסה נוספת המכונה "מצב זיכרון התא" ומסומנת על ידי c_{t-1} . בנוסף לכך מצב התא מחושב בצורה מורכבת יותר מאשר תא RNN.

הצמד $[x_t, h_{t-1}]$ נכנס אל תא ה-LSTM ולאחר מכן עובר דרך מספר "שערים" (כל הפעולות מבוצעות בנפרד על כל אחד מהרכיבים):

1. שער השכחה (ft) - שער השכחה הוא שכבה סיגמואידית שמחליטה איזה מידע ממצב התא יש לזרוק או לזכור. הוא לוקח כקלט את המצב הנסתר הקודם (h_{t-1}) ואת הקלט הנוכחי (x_t) ומייצר ערך "שכחה" (ft) הנע בין 0 ל-1 עבור כל אלמנט במצב התא. לאחר מכן מוכפל ערך זה במצב זיכרון התא c_{t-1} .
2. שער "קלט" - שער הקלט קובע איזה מידע חדש יש להוסיף למצב התא. שער זה מורכב משני חלקים: שכבת סיגמואידית ושכבת tanh (אשר אמונה על יצירת וקטור של ערכים חדשים). את תוצאות שער הפלט מוסיפים לערכי ווקטור c_{t-1} .
3. שער פלט - שער הפלט מחליט מה צריך להיות המצב הנסתר הבא (h_t). הוא מורכב משכבת סיגמואידית המייצרת ערך שער פלט (ot) ושכבת tanh המייצרת את המצב הנסתר הבא.



איור 17 - ארכיטקטורת LSTM [8].

באופן זה מקבלים את האינדקס h_t שאחראי כמו בתא ה-RNN על הזיכרון המועבר בטווח הקצר, ובנוסף מקבלים את האינדקס c_t האחראי על כלל הזיכרון שנאסף בעבר. הארכיטקטורה של הרכיב מאפשרת להתייחס לשלל אלמנטים הנוגעים לזיכרון – ניתן לשכוח חלקים לא רלוונטיים ברצף הקלט ולהדגיש חלקים משמעותיים ברצף הקלט.

LSTMs מצטיינים בלכידת תלות ארוכת טווח בנתונים עוקבים. מצב התא, הנשלט על ידי שערי השכחה, הקלט והפלט, מאפשר לרשת לאחסן ולאחזר מידע באופן סלקטיבי לאורך שלבי זמן מרובים. בנוסף LSTM מצליחים להתמודד עם בעיית הגרדיאנט הנעלם בצורה מוצלחת יותר. מנגנוני השער מאפשרים הפצה אפקטיבית לאחור של אותות שגיאה, ומאפשרים לרשת ללמוד ולהסתגל על פני רצפים מורחבים [24].

שער השכחה מאפשר ל-LSTMs לשמור או להשליך מידע באופן סלקטיבי, מה שהופך אותם לחזקים בתרחישים שבהם יש לסנן פרטים לא רלוונטיים. ישנם יישומים רבים של רשתות LSTM כגון עיבוד שפה טבעית (NLP) כגון תרגום שפה, ניתוח סנטימנטים ויצירת טקסט, כאשר הבנת הקשר על פני משפטים ארוכים היא חיונית או למשל בניתוח סדרות זמן, בתחזיות פיננסיות, חיזוי מזג אוויר ותחומים אחרים תלויי זמן [3].

לסיכום, רשתות זיכרון ארוך טווח קצר הן כלי יעיל בנוף הלמידה העמוקה עבור משימות הכוללות נתונים עוקבים. היכולת שלהם ללכוד ולשמור על תלות ארוכת טווח מיקמה אותם כטכנולוגיה בסיסית ביישומים שונים, ומחקר מתמשך ממשיך לשכלל ולהרחיב את היכולות שלהם.

יחד עם זאת לרשתות הזיכרון יש גם חסרונות. על אף המבנה המורכב יותר של רשתות זיכרון מסוג LSTM גם רשתות אלו מתקשות לבצע היקשים על רצפים ארוכים מאוד. הסיבה לכך תמונה בעובדה שרשתות אלו נדרשות לעבד את המידע המצוי בקלט מסוים, מתחילתו ועד סופו. ההיקשים שהרשת לומדת נובעים מעיבוד כולל של כלל המרכיבים בקלט וזאת תוך כדי לימוד הסדר הכרונולוגי של הקלט. במילים אחרות, יש חשיבות מכרעת בין תוצאות הקלט ובין סדר האיברים בקלט. עובדה זו גורעת מיכולות מיצוי המידע של רשתות נשנות. שכן לעיתים כדי להפיק את מירב המידע מקלט יש לבחון את עוצמת החיבורים שבין אבני הבניין שלו ולא רק בצורה כרונולוגית. בפרק הבא אסקור את מנגנון הקשב אשר נותן מענה לבעיה זו.

3.4 מנגנון הקשב וטרנספורמרים

טרנספורמר (Transformer) הוא מודל למידה עמוקה המבוסס על מנגנון הקשב (Attention). המודל הוצג לראשונה ע"י [7] וזכה להשפעה רבה על עולם הלמידה העמוקה ככלל ובמיוחד בכל הנוגע לתהליך עיבוד רצפי מידע.

כפי שסקרתי בפרק הקודם, עיבוד רצפי מידע בוצע בעיקר ע"י רשתות נשנות בעלות תאים נשנים מסוגים שונים. מנגנון הקשב ככלל ומודל הטרנספורמר הציגו גישה שונה לעיבוד רצפי מידע. מנגנון הקשב מאפשר לרשת הנוירונית להתמקד בחלקים שונים של רצף הקלט בעת ביצוע תחזיות, לשקלל את התלויות וההקשרים בין חלקי הרצף ועל ידי כך לבצע תחזיות וסיווגים מדויקים יותר.

בניגוד לרשתות נשנות המעבדות רצפי מידע על ידי שימוש בתאי זיכרון השומרים בתוכם מידע לגבי רצף המידע שעובד עד נקודת זמן t מנגנון הקשב פועל בצורה שונה. המנגנון ממפה את ההקשרים הקיימים בין נתוני רצף הקלט ומתמקד בהקשרים ותלויות חזקים בין מקטעים בתוך הרצף. מנגנון הקשב מאפשר לכל פיסת מידע ברצף הנתונים "להתייחס" לכל פיסות המידע האחרות, תוך מתן משקלים שונים לקשר בינן לבין עצמן. מדידת ההקשרים מבוצעת "במכה אחת" ומנגנון הקשב איננו שומר ייצוג "זיכרון" לגבי רצף הקלט. בכך מנגנון הקשב מאפשר ללמוד ייצוגים ותלויות שונות הקיימות בתוך הקלט עצמו ובין אבני הבניין השונות המרכיבות אותו. פלט מנגנון הקשב הוא ייצוג של ההקשרים בין כלל אבני הבניין המרכיבים את הקלט ובכך הפלט מכיל מידע רב בהרבה מאשר הפלט של רשתות נשנות [14].

3.4.1 תיאור סכמטי מנגנון (שכבת) הקשב

בכדי לדון באופן פעולת שכבת הקשב נחدد את ההבדל בין מנגנון זה לבין תא רשת נשנית דוגמת LSTM. הפלט של תא רשת נשנית הוא כאמור ווקטור דחוס של נתונים שנקרא ווקטור ההקשר. ווקטור זה אמור להכיל בתוכו את כל ההקשרים החבויים בתוך הקלט. הבעיה מתעוררת כאשר קלט הרצף הוא קלט ארוך המכיל תלויות והקשרים רבים. במצב זה ווקטור ההקשר איננו מצליח להכיל בתוכו את כלל ההקשרים והתלויות החבויים בתוך הקלט [8]. כיצד אם כן שכבת הקשב נותנת מענה לבעיה זו. שכבת הקשב בוחנת כל אחד מחלקיו של הקלט, כלומר את כל אחת מאבני הבניין המרכיבות את הקלט.

נדגים את פעולת מנגנון הקשב באמצעות דוגמה מילולית. נניח שהקלט של מנגנון הקשב הוא משפט תקני בשפה העברית. מנגנון הקשב מעבד את כל אחת מהמילים המרכיבות את המשפט. לאחר מכן המנגנון מעניק ציון לכל אחת מהמילים הבונות את המשפט.

המנגנון מחשב את ההקשר בין הציונים השונים שניתנו לכל אחת מהמילים בכדי לנסות ולזהות הקשרים ותלויות בין המילים.

ניקח לדוגמה את המשפט **"הילד לא חצה את הכביש בגלל שהוא היה עייף מידי"**.

בשונה משכבת ה-LSTM אשר הייתה מנתחת את המשפט מתחילתו ועד סופו, כלומר, שכבת LSTM הייתה מעבדת את המילה הראשונה במשפט, מייצגת אותה בצורה ווקטורית ולאחר מכן מעבדת את המילה השניה במשפט וכך הלאה בניסיון לדלות את המידע מרצף המילים על פי הסדר שלהן במשפט, מנגנון הקשב מבצע פעולה שונה.

המנגנון ייתן ציון לכל אחת מהמילים המרכיבות את המשפט וינסה למפות את הקשר ביניהן כדי למצות את המידע וההקשרים הקיימים במשפט. ניתן לדוגמה לראות שהמילה הראשונה "הילד" קשורה בצורה משמעותית יחסית למילה "הוא", שהרי הכוונה במילה "הוא" היא לילד.

בשונה מתא LSTM אשר היה צריך לעבד מידע רב בין שתי המילים הללו (בעקבות מיקומן הרחוק בתוך המשפט הנתון) מנגנון הקשב ייתן ציון גבוה לקשר בין שתי המילים כיוון שהן בעלות זיקה משותפת רבה. על ידי כך מאפשרת שכבת הקשב ניתוח מבוסס הקשר פנימי של קלטים מבוססים רצפי זמן.

פלט שכבת הקשב הוא אם כן מטריצה של ציונים עבור כל זוג מילים במשפט הקלט. ככול שהציון בין זוג מילים גובה יותר כך הקשר ביניהן משמעותי יותר. פלט זה מאפשר לרשת הנורונים ללמוד על הקשרים בשכבת הקלט. מנגנון קשב דינמי זה מאפשר למודל ללכוד תלות ויחסים ארוכי טווח בתוך רצף הקלט, מה שהופך אותו ליעיל ביותר עבור משימות הדורשות הבנה של הקשר ותלות גלובלית [8].

3.4.2 תיאור פורמלי של מנגנון הקשב

נדגים את פעולת מנגנון הקשב באופן פורמלי. מנגנון הקשב יוצר שלוש מטריצות של מקדמים עבור סדרת הקלט. המטריצות נקראות Query, Key, Value והן נוצרות על ידי הכפלה של מטריצת המשקלים עם איבר הקלט. באמצעות מטריצות אלו מחשבים את ה attention score. ערך זה מביע את החשיבות של כל זוג ערכים (שאלת Q ומפתח K)

ומתבטא בערך V . ככל שהערך V עבור זוג Q ו- K גדול יותר ניתן להסיק שלמפתח K יש רלוונטיות גבוהה יותר עבור השאילתא Q . אציג להלן הסבר כיצד מחושב הערך בפירוט.

$$\text{Attention}(\text{Query}, \text{Key}, \text{Value}) = \text{SoftMax}\left(\frac{Q \cdot K}{\sqrt{d_k}}\right) \cdot V$$

משוואה 7 - אופן חישוב attention score [14].

נבחן את המשוואה באיור 22. עבור סדרת קלט X מחושבות שלוש מטריצות כאשר כל איבר בסדרה המקורית x_i יוצר שורה בכל אחת מהמטריצות. כאשר לוקחים את השורה $q_i = Qx_i$ ומכפילים אותה בשורות המטריצה K מקבלים ווקטור חדש שכל איבר j בווקטור הנ"ל מציין את עוצמת הקשר בין האינדקסים i, j בווקטור הקלט.

ביצוע ההכפלה הנ"ל עבור כל אחד מאיברי הקלט יוצר מטריצה שבה כל שורה מייצגת את הקשר בין איבר i לשאר האיברים בסדרת הקלט. ההכפלה הזו היא בעצם כפל המטריצות QK . כדי לשמור על יציבות הגרדיאנט מחלקים בשורש מימד ה-embedding ומבצעים נרמול באמצעות שכבת SOFTMAX.

אם נסמן כל אינדקס במטריצה w_{ij} נוכל לקבל אותו ישירות בנוסחה הבאה:

$$w_{ij} = \text{SoftMax}\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right) = \frac{\exp\left(\frac{q_i^T k_j}{\sqrt{d_k}}\right)}{\sum_{s=1}^n \exp\left(\frac{q_i^T k_s}{\sqrt{d_k}}\right)}$$

משוואה 8 - האינדקס w_{ij} [14].

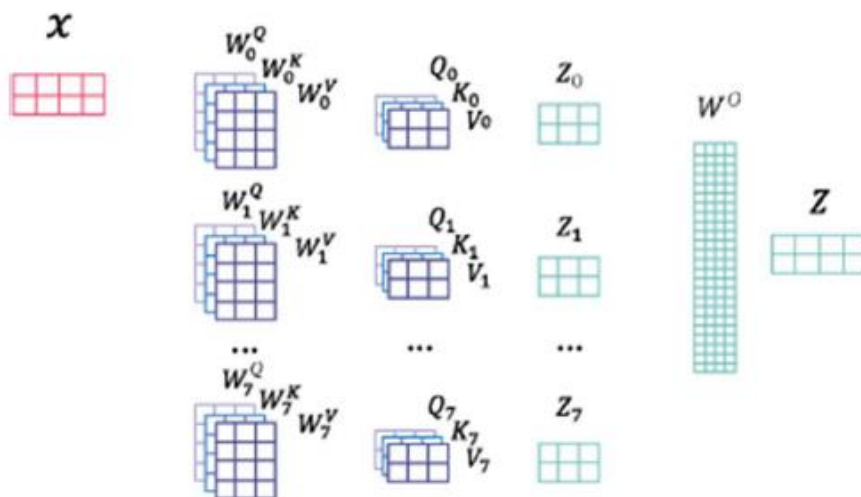
לאחר שאנו יודעים למצוא את האינדקס w_{ij} בונים ייצוג חדש לקלט המקורי ע"י הכפלה במטריצה V :

$$z_i = \sum_{j=1}^n w_{ij} v_j = \frac{\sum_{j=1}^n \exp(q_i^T k_j) v_j}{\sum_{s=1}^n \exp(q_i^T k_s)}$$

משוואה 9 - ייצוג חדש של איברי סדרת הקלט במטריצה z [14].

הסדרה המתקבלת היא ייצוג חדש של סדרת הקלט המקורית כאשר כל איבר z_i בסדרה החדשה מייצג איבר בסדרה המקורית עם כלל ההקשרים בינו לבין שאר איברי הסדרה.

ניתן להשתמש במנגנון הקשב מספר פעמים במקביל וליצור מספר מטריצות מטרה עצמאיות. שלישיית מטריצות עצמאית QVK נקראת גם ראש קשב. חיבור של מספר ראשי קשב נקרא מנגנון קשב רב ראשי. באופן זה לכל איבר קלט x_i יש כמה ייצוגים שונים z_i אותם ניתן להכפיל במטריצת משקלים w_0 ולקבל ייצוג משוקלל של אותו איבר.



איור 18 - מנגנון קשב רב ראשי [14].

3.4.3 טרנספורמרים

מנגנון הקשב הוצג לראשונה כחלק מארכיטקטורה הנקראת טרנספורמר. טרנספורמר הוא שם שניתן למודל המבוסס על ארכיטקטורת מקודד ומפרש שיועד לתת מענה מתקדם לעיבוד רצפי מידע. אבני הבניין של הטרנספורמר מחולקים אפוא לשניים – מקודד ומפרש.

תפקידו של המקודד הוא למקסם את המידע המופק מעיבוד הקלט. המקודד מקבל סדרת קלט x (לרוב אחרי שעברה תהליך עיבוד אותו אציג בהמשך) ומנסה לחלץ את המידע הטמון בסדרת הנתונים ולצמצם את השפעתם של רעשים. המקודד בנוי משכבות מרובות הנערמות זו על זו וזהות במבנה שלהן, כל שכבה נקראת גם תא קידוד. בכל תא קידוד מחלוק לשני חלקים מרכזיים החלק הראשון של תא הקידוד הוא מנגנון הקשב מרובה ראשים והחלק השני של התא הוא רשת נוירונים.

המידע זורם אל תא הקידוד ראשית אל שכבת הקשב שם מבוצע עיבוד של הקלט. לאחר מכן הפלט של שכבת הקשב משורשר בחזרה אל הדאטה הגולמי שהוזן לתא המקודד באמצעות מנגנון שיויר. המידע המשורשר מוזרם לתוך שכבת נרמול.

לאחר מכן פלט שכבת הנרמול מוזן לחלקו השני של תא המקודד שהוא כאמור רשת נוירונים FC המנורמלת בסופה אף היא. במודל הטרנספורמר ניתן לחבר מספר תאי קידוד כאשר הפלט של תא קידוד אחד הוא הקלט של תא הקידוד הבא. טרנספורמרים שונים מובחנים אחד מהשני בכמות תאי הקידוד המכילים אותם.

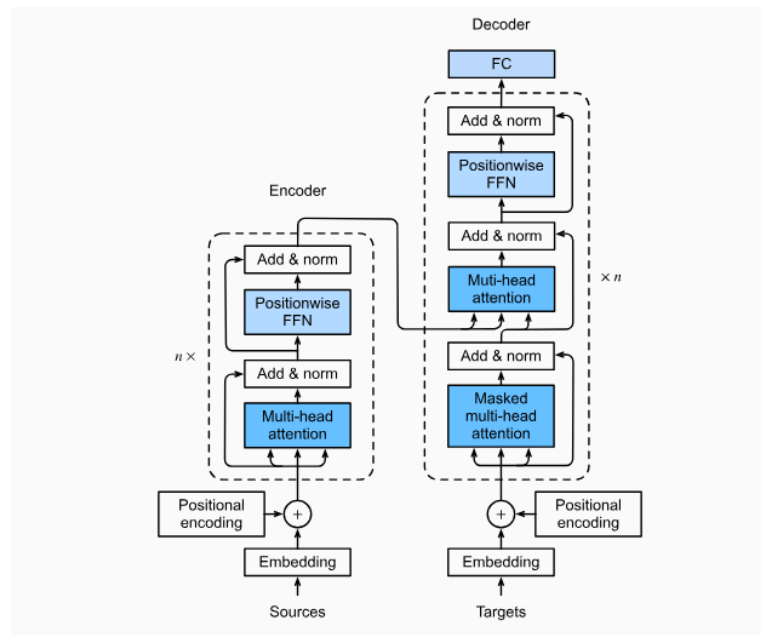
המפרש – חלקו השני של הטרנספורמר בנוי בצורה דומה למקודד אך הוא בעל שני הבדלים משמעותיים. ראשית הקלט של המפרש הם איברי הפלט של המקודד, כלומר רק בסוף תהליך הקידוד על כל תאי הקידוד של המקודד המידע מוזן לרכיב המפרש.

שנית בתחילתו של המפרש ישנו רכיב נוסף הנקרא Masked self-attention. שכבה זו מקבלת את כל רכיבי הפלט שהופקו ובאמצעותה ינסה המפרש "לנחש" את האיבר הבא של סדרת הפלט.

בשלב הראשון משתמש המפרש בשכבת self-attention על איברי הסדרה שהתקבלו עד כה ולומד ייצוג חדש שלהם. לאחר השכבה הראשונה יש שכבת מנגנון קשב רב ראשי הנקרא Encoder-Decoder Attention, שהוא שילוב של המקודד והמפרש – המטריצות Key, Value נלקחות מהמקודד והמטריצה Q מגיעה מהמפרש.

כאשר מבצעים את מכפלת המטריצות KQ לא מחפשים דמיון בין האיברים באותה הסדרה אלא מחפשים דמיון בין האיברים של סדרת הפלט (כפי שיוצגו על ידי המקודד) לבין איברי סדרת הקלט (כפי שיוצגו ע"י המפרש באמצעות שכבת Masked self-attention).

שימוש זה מאוד דומה למנגנון שכבת הקשב המקורי אך לא נעשה בו שימוש ברכיבי זיכרון, כלומר המכפלה KQ מייצרת מטריצה של משקלים שכל איבר ואיבר במטריצה מציין מה היחס בין איבר בסדרה המקורית ובין איבר בסדרת הפלט את המטריצה הזו מכפילים במטריצה V המגיעה כאמור מהמקודד וכך מקבלים ווקטור שהוא ייצוג חדש של איבר הפלט הבא. איבר זה עובר אף הוא בשכבות FC SOFTMAX.



איור 19 - מודל הטרנספורמר [7]

רכיב נוסף הנמצא בכניסה למפרש ולמקודד הוא רכיב המבצע "קידוד מיקום". לטרנספורמר אין מידע מובנה על סדר הנתונים בתוך רצף הקלט. מנגנון הקשב כאמור איננו בעל תא זיכרון השומר מידע על רצף נתוני הקלט. יחד עם זאת מובן כי כאשר עוסקים בקידוד רצפים המשתנים לאורך זמן ישנה חשיבות גבוהה להבנת המיקום של הנתונים לאורך הרצף, ומיקום הנתונים יכול לעזור ולהבין את ההקשרים ביניהם ואת משמעותם.

אי לכך במודל הטרנספורמר משתמשים בקידוד מיקום מוכנס כדי לאפשר למודל להשתמש במיקום של הנתונים על פני רצף הקלט. במילים אחרות כל פיסת מידע ברצף הנתונים מקבל ציון או ניקוד בהתאם למיקומו היחסי ברצף המידע. תוספת זו מבטיחה שהמודל יכול להבחין בין מיקומם של נתונים שונים לאורך הרצף ועובדה זו חיונית למשימות שבהן סדר הנתונים חיוני לפענוח רצף המידע, כגון תרגום שפה [7].

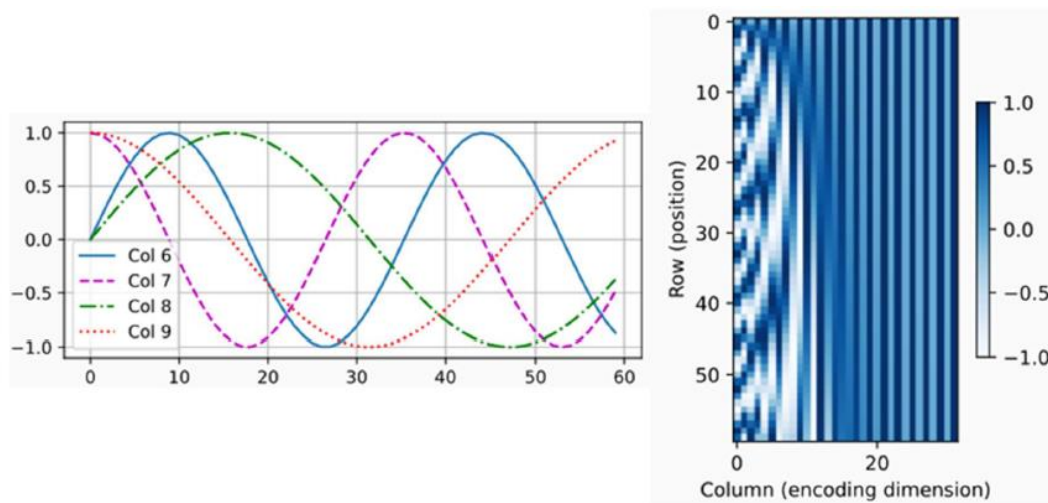
כיצד עובד קידוד מיקום – בצורת קידוד זו מוסיפים לכל איבר קלט פיסת מידע לגבי מיקומו בסדרה, המבוסס על פונקציות טריגונומטריות. נדגים את פעולת הרכיב בצורה פשוטה באמצעות ייצוג בינארי של מספרים. אם ניקח רצף של מספרים בינאריים נוכל לראות שכול שלמספר יש ערך מספרי גבוה יותר כך הוא משתנה בתדירות נמוכה יותר. תדירות השינוי של ספרות המספר הבינארי יכולים אי לכך להעיד על מיקומו ברצף.

0:	0	0	0	0
1:	0	0	0	1
2:	0	0	1	0
3:	0	0	1	0
4:	0	1	0	0
5:	0	1	0	1
6:	0	1	1	0
7:	0	1	1	1
8:	1	0	0	0
9:	1	0	0	1
10:	1	0	1	0
11:	1	0	1	1
12:	1	1	0	0
13:	1	1	0	1
14:	1	1	1	0
15:	1	1	1	1

איור 20 - ייצוג בינארי של מספרים, ה-MSB משתנה בתדירות שונה לעומת ה-LSB [14].

באיור מעלה ניתן להבחין כי תדירות השינוי של הספרה הימנית ביותר במספר הבינארי משתנה בקצב גבוה בהרבה מהספרה השמאלית ביותר. ניתוח ושימוש בתדירות השינוי של כל ספרה יכול להעיד על גודל המספר או על מיקומו בסדרת המספרים.

כיוון שאנו לא עוסקים במספרים שלמים, הייצוג הבינארי לא יהיה מתאים לשימוש. ולכן ניתן להשתמש בפונקציות טריגונומטריות שהן רציפות ובעלות תדירות משתנה. זהו למעשה הרכיב שאנו מוסיפים לאיברי הקלט טרום הזנתם לרכיבי הטרנספורמר – רצף של ערכי פונקציות טריגונומטריות עם תדירות הולכת וקטנה. כך לכל רכיב בקלט מקבל ציון או ערך המעיד על מיקומו בסדרה.



איור 21 - המחשה לקצב השינוי של פונקציות טריגונומטריות בעלות תדרים משתנים [14].

3.5 רשתות גנרטיביות - GAN

הכלים שהוצגו לעיל עוסקים בסיווג של דוגמאות קלט למחלקות ידועות מראש. בלמידה עמוקה קיימים כלים נוספים שמטרתם העיקרית היא איננה סיווג בין נתוני קלט לתייגים אלא רשתות שמטרתם ייצור נתונים סינטיים אשר דומים עד כדי זהות מוחלטת (מבחינה הסתברותית) לנתונים גולמיים שנדגמו מהעולם האמיתי. רשתות נוירונים גנרטיביות הן רשתות המסוגלות לייצר מידע סינטי מסוג זה. רשתות יריבות גנרטיביות (GANs) הן סוג של רשתות נוירונים המאפשרת למודלים ליצור תוכן ריאליסטי ואיכותי, החל מתמונות וכלה במוזיקה וסוגי נתונים נוספים. רשתות GAN התפתחו לכלים רבי עוצמה ונמצאים בשימוש עבור יישומים בתחומים שונים כדוגמת יישומים מתחום עיבוד השפה הטבעית ומודלי שפה (chat GPT, וכלי שפה מתחרים אחרים) ובתחום הראייה הממוחשבת ויצירת אומנות גרפית לשימושים שונים [15].

רשתות אלו נמצאות בשימוש בתחומים רבים. הן במשימות עיבוד שפה טבעית ובמשימות הנוגעות לראייה ממוחשבת כגון ייצור סינטי של תמונות וסרטונים. המוטיבציה להשתמש בGAN נובעת בין היתר מהיכולות של רשתות אלו לתרום לתהליכים רבים בתחום המחקר. לדוגמה במידה וברשותנו מערך נתונים קטן ומתויג וברצוננו להגדיל את מערך הנתונים, נוכל לבנות מודל גנרטיבי. המודל יזהה את המאפיינים, התכונות והייצוגים המרכזיים של מערך נתונים הקטן שברשותנו ויוכל להגדיל את מערך הנתונים בדוגמאות חדשות בצורה סינטיטית. דוגמאות סינטיטיות אלו יהיו בעלות התפלגות סטטיסטית קרובה מאוד לדוגמאות האמת. ובכך נוכל לדוגמה להגדיל את קבוצת המדגם באמצעות דאטה סינטיטי.

אדגים את הקביעה הנ"ל באמצעות דוגמה מוחשית. בהינתן מאגר גדול של תצלומים של פרצופים, נרצה להיות מסוגלים ליצור תמונה ריאליסטית חדשה שנראית כאילו נדגמה ממאגר התמונות הגולמי ובפרט שהתפלגות התמונה זהה להתפלגות מרחב המדגם. כך נוכל למשל להרחיב את מאגר המידע שלנו באמצעות דוגמאות מתויגות סינטיטיות או לחלופין לספק כפלט פרצוף חדש של אדם שנראה ריאליסטי לחלוטין [8].

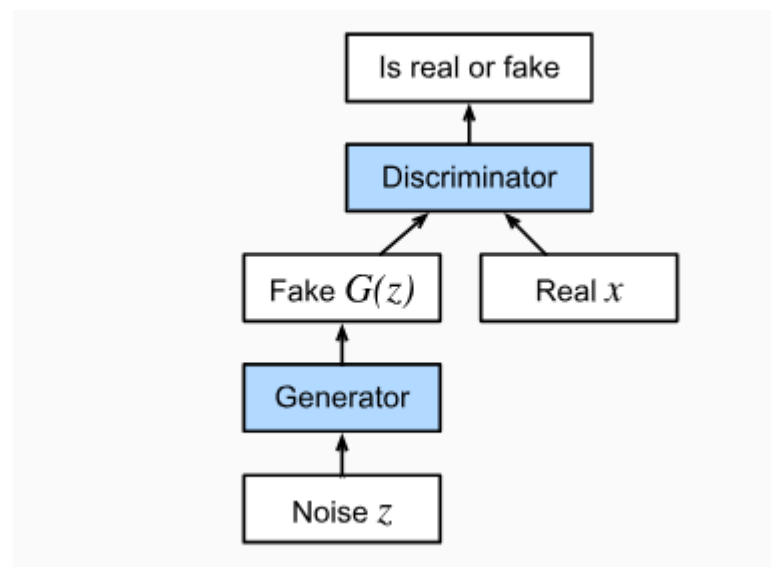
3.5.1 מבנה רשת גנרטיבית

הרעיון מאחורי רשת גנרטיבית הוא למנף את כוחם של מודלים מפלים (מודלים העוסקים בסיווג) כדי ליצור מודלים גנרטיביים. הרעיון מסתמך על ההנחה שמחולל נתונים הוא טוב אם אנו איננו יכולים להבדיל בין נתונים מלאכותיים שיוצרו על ידי המחולל הסינטיטי ובין נתונים אמיתיים. כלומר, מחולל נתונים יהיה מחולל נתונים טוב אם הוא מייצר דוגמאות סינטיטיות בעלות התפלגות נתונים סטטיסטיים הזוהות לדוגמאות המקור.

הכוח של רשתות GAN נובע מהשימוש השונה שעושים במודלים מפלים כחלק מתהליך הלימוד והעיבוד שלהם. במילים אחרות, במקום לבנות מודל שמטרתו לזהות ששני מערכי נתונים לא הגיעו מאותה התפלגות סטטיסטית – לדוגמה מודל שמזהה תמונות של כלבים ומבדיל אותם מתמונות של חתולים, משתמשים במודל הסיווג כדי לספק אותות אימון למודל המחולל נתונים סינטיים חדשים ובעזרת אותות אלו מנסים לשפר את תהליך ייצור הדוגמאות הסינטיים עוד ועוד. התהליך זה מאפשר לנו לשפר את מחולל הנתונים עד שהוא יוצר דוגמאות כמעט זהות לנתונים אמיתיים [8].

בליבת ה-GANs שתי רשתות עצביות מופרדות הנקראות המחולל (גנרטור) והמפלה (דיסקרימינרטור). אימון הרשת כולל העברת נתונים ושיפור מתמיד בין שתי תתי הרשתות הללו. מטרת המחולל היא ליצור נתונים סינטיים, תוך ניסיון לחקות את התפלגות הנתונים האמיתיים. אם בתמונות עסקינן אזי המחולל נדרש לייצר תמונות סינטיטיות הנראות דומות ככול הניתן לתמונות המקור.

הארכיטקטורה הבסיסית של מודל GAN מוצג באיור הבא:



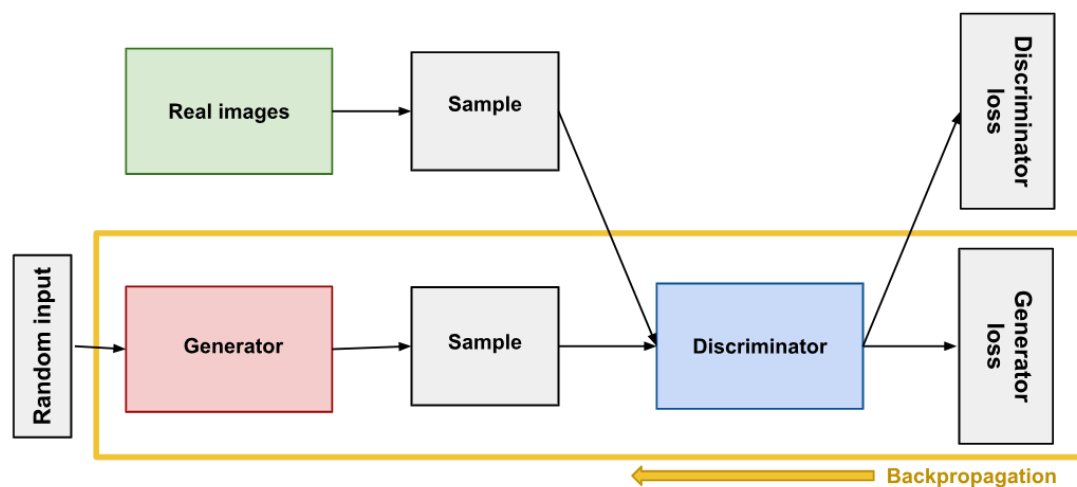
איור 22 - ארכיטקטורת מודל GAN [8].

3.5.2 אימון הרשת

המפלה מנסה להעריך אם נתוני הקלט שהוזנו לו הגיעו ממקור הנתונים האמיתי או שנוצרו על ידי המחולל. ניתן להסתכל על תהליך האימון בצורה הבאה – המחולל מנסה ליצור תמונות אמיתיות ככל הניתן בכדי "לרמות" את המפלה. המפלה מנסה לקבוע בכל זמן האם התמונה שקיבל היא תמונה אמיתית או תמונה שיוצרה על ידי המחולל. ציון פונקציית ההפסד

של המחולל מעדכנים את משקלי שתי הרשתות במטרה לשפר את שתיהן ובכך לשדרג את יכולות המחולל והמפלה [8].

המחולל דוגם רעש אקראי כקלט ובאמצעות רשת נוירונים עמוקה הופך את הרעש לנתונים כך שבאופן אידיאלי אמורים להיות זהים לחלוטין לנתונים אמיתיים. לאחר מכן מוזרמים הנתונים הסינטטיים שיוצרו על ידי המחולל אל רשת המפלה אשר בתורו מנסה לסווג את הנתונים לאמיתיים או סינטטיים. ציון פונקציית ההפסד של המפלה משמש בתורו את המחולל – ככול שהמחולל ייצר דוגמה "גרוע" יותר כך הוא מתקן את משקולותיו בכדי לטייב ולדייק את ייצור הדוגמה הבאה.

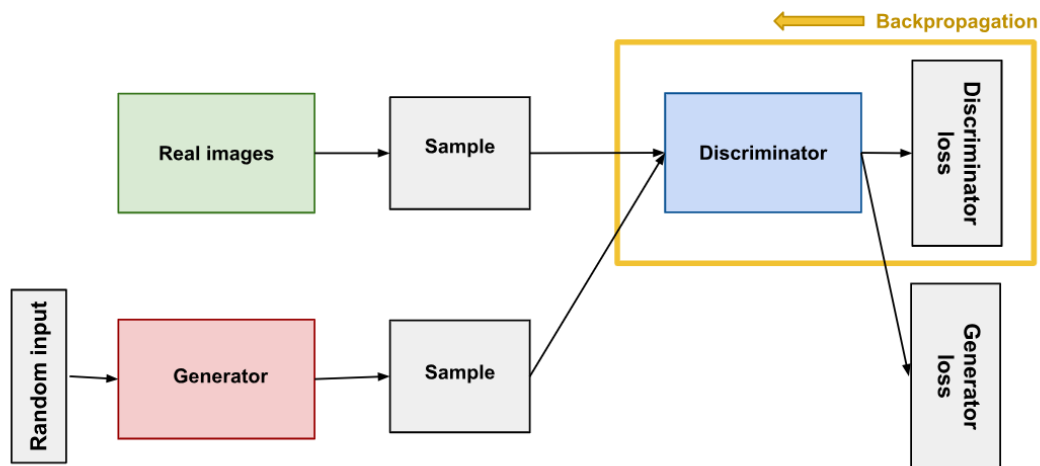


איור 23 - תהליך למידת המחולל [8].

המפלה, שהוא כאמור רשת עצבית נוספת, שואף להצליח להבחין בין נתונים אמיתיים למידע שנוצר על ידי המחולל. המפלה מאומן להיות יותר ויותר מדויק במשימת אבחנה זו. תהליך האימון המשותף למחולל ולמפלה מביא לשיווי משקל דינמי שבו המחולל שואף ליצור נתונים מציאותיים מספיק כדי להטעות את המפלה. נתוני האימון של המפלה מגיעים משני מקורות:

1. מופעי נתונים אמיתיים, כגון תמונות אמיתיות של אנשים. המפלה משתמש במקרים אלה כדוגמאות חיוביות במהלך האימון.
2. מופעי נתונים מזויפים שנוצרו על ידי המחולל. המפלה משתמש במקרים אלה כדוגמאות שליליות במהלך האימון.

במהלך האימון של המפלה, המחולל אינו מתאמן. משקליו נשארים קבועים בזמן שהוא מייצר דוגמאות לאיבחון. אימון המפלה מתבצע בצורה הבאה - המפלה מסווג בן נתונים אמיתיים ובן נתונים מזויפים מהמחולל. פונקציית ההפסד של המפלה מענישה את המפלה על סיווג שגוי של מקרה אמיתי כמזויף או מקרה מזויף כאמיתי. המפלה מעדכן את משקולותיו באמצעות פעפוע לאחור לאור ערכי פונקציית ההפסד [15].



איור 24 - תהליך למידת המפלה [8].

מכיוון ש-GAN מכיל שתי רשתות נוירונים המאומנות בנפרד, אלגוריתם האימון שלו חייב לטפל בשני קשיים מרכזיים::

1. תהליך האימון חייב לתמוך בין שני סוגים שונים של אימון (מחולל ומפלה).
2. קושי לזהות התכנסות GAN.

אי לכך אימון GAN מתקיים באיטרציות מתחלפות:

1. המחולל מתאמן במשך איטרציה אחת או יותר.
2. המפלה מתאמן במשך איטרציה אחת או יותר.
3. חוזרים על שלבים 1 ו-2 כדי להמשיך לאמן את רשתות עד להגעה לרמה הנדרשת.

משקולות המחולל נשארות קבועות בשלב האימון של המפלה. בעוד שבמהלך אימונו של מפלה הוא מנסה ללמוד כיצד להבחין בין נתונים אמיתיים לבין מזויפים, עליו ללמוד כיצד לזהות את הפגמים של המחולל, במילים אחרות עליו ללמוד את הדפוסים הכושלים של המחולל בכדי לעמוד על השוני בין המידע הסינטטי שייצר לבין המידע האמיתי. באופן דומה, אנו שומרים על משקולות המפלה קבועות במהלך שלב האימון של המחולל.

3.5.3 פונקציית ההפסד של GAN

המפלה הוא למעשה מסווג אשר הפלט שלו הוא מידת ההסתברות שהקלט שהוכנס אליו הינו דוגמה אמיתית מהעולם האמיתי. נסמן את ההסתברות הזו ב $D(x)$. בכדי לאמן את המפלה נרצה להשיג שתי מטרות, ראשית נרצה ש $D(x)$ יהי מקסימלי עבור דוגמה אמיתית, ובמקביל שערך זה יהיה מינימלי עבור דוגמה סינטטית שנוצרה על ידי המחולל.

בצורה דומה נרצה לאמן את המחולל. מטרתו של המחולל היא להערים על המפלה בכך שייתן ציונים גבוהים לדוגמאות סינטטיות. כדי לייצג את פער זה נבנה פונקציית מחיר בעלת שני איברים:

$$V(D, G) = \min_G \max_D \mathbb{E}_{x \sim Data} \log D(x) + \mathbb{E}_{z \sim Noise} \log(1 - D(G(z)))$$

משוואה 10 - פונקציית המחיר של במודל GAN [14].

כאשר X מייצג דוגמה אמיתי ממאגר המידע $G(Z)$ מייצג תמונה סינטטית שיוצרה על ידי המחולל. $D(x)$ מייצג את ערך הניבוי של המפלה עבור דוגמה אמיתית, ו $D(G(Z))$ מייצג את ערך המפלה לגבי תמונה סינטטית.

המפלה מעוניין למקסם את פונקציית המחיר כך שהערך של $D(x)$ יהיה קרוב ככול הניתן ל-1 ו $D(G(Z))$ יהיה כמה שיותר קרוב ל-0, המחולל לעומת זאת מעוניין שהביטוי $D(G(Z))$ יהיה קרוב ככול הניתן ל-1, כלומר שמחולל יחשוב ש $G(Z)$ הוא דאטה אמיתי.

האימון כאמור מתבצע בצורה איטרטיבית כך שבכל פעם מקבעים את G או את D .

אם מקבעים את המחולל, אנו מאמנים למעשה מסווג בינארי כאשר מחפשים את האופטימום התלוי בווקטור הפרמטרים θ_d :

$$\max_{\phi_d} \mathbb{E}_{x \sim Data} \log D_{\phi_d}(x) + \mathbb{E}_{z \sim Noise} \log(1 - D_{\phi_d}(G_{\theta_g}(z)))$$

משוואה 11 - [14].

לעומת זאת, במידה ונקבע את המפלה, ניתן להתעלם מהאיבר הראשון במשוואה, כאשר מטרתנו היא לאמן את המחולל באמצעותו של המפלה במצבו הנוכחי, ולכן נשאר רק עם הביטוי השני:

$$\min_{\theta_g} \mathbb{E}_{z \sim Noise} \log(1 - D_{\phi_d}(G_{\theta_g}(z)))$$

משוואה 12 - [14].

מטרת המפלה היא למזער את ערך פונקציית ההפסד הנ"ל ע"י עדכון המשקולות שלו בכדי לייצר ערכים הנראים זהים לתמונות אמיתיות.

3.5.4 מבחן והערכת ביצועים ברשתות גנרטיביות

רשתות גנרטיביות אינן עוסקות בסיווג מידע אלא בייצור של מידע חדש וסינטטי ולכן תהליך המבחן של רשתות אלו שונה מתהליך המבחן של רשתות נוירונים אחרות. מטרת הרשת הגנרטיבית היא אפוא לייצר דאטה סינטטי שאמור להיות זהה ככול הניתן לדאטה מהעולם שנדגם בעולם האמיתי.

אי לכך פעמים רבות כלי הערכה הטוב ביותר הוא בדיקה וויזואלית של משתמשים אנושיים באם מדובר בדאטה המאפשר זאת. אם למשל הרשת הגנרטיבית אמורה ליצור תמונות סינטטיות של בעלי חיים, בחינה אנושית של התמונות תוכל לתת מענה לבדיקת ביצועי הרשת. בוחנים אנושיים יכולים להעריך את ביצועי הרשת ולבדוק את מידת האותנטיות והראליסטיות של פלט המערכת. צורת הערכה זו יכולה להיות אומנם אפקטיבית שהרי מטרת המערכת היא לייצר דאטה סינטטי הנראה כמו דאטה אמיתי אך שיטה זו גם קשה להשוואה וכימות אמפירי.

שיטות הערכת ביצועים אחרות מתבססות על מטריקות הערכה שונות כגון מטריקת FID. מטריקת FID (Frechet Inception Distance) היא מטריקה שיעודה הערכת ביצועים עבור מודלים גנרטיביים ובמיוחד עבור תמונות חזותיות. המטריקה מודדת את רמת הדמיון בין התפלגות ערכי הדוגמה הסינטטית יחסית לערכי דוגמה [14].

חישוב מטריקת FID כולל מספר שלבים. בשלב הראשון נדרש לחלץ פיצ'רים (מאפיינים) מייצגים מדוגמאות אמיתיות ומדוגמאות סינטטיות שנוצרו על ידי המודל באמצעות מודל נוסף. מאפיינים אלו מדוגמאות האמת ומדוגמאות הסינטטיות אמורות לייצג מידע "סמנטי" ממימד גבוה על הדוגמאות. לאחר דגימת המאפיינים המטריקה מחשבת את הממוצע והשונויות עבור הדוגמאות הסינטטיות והאמיתיות, נתונים אלו מייצגים בצורה מתומצתת את התפלגות המאפיינים במאגרי הנתונים. לאחר מכן המטריקה מחשבת את מרחק פרצ'ט (Fréchet) בין ההתפלגויות הגאוסיאניות המוגדרות על ידי הממוצע והשונויות של הדוגמאות הסינטטיות והאמיתיות. ככול שערך המטריקה נמוך יותר כך הדימיון בין ההתפלגויות הוא טוב יותר [14].

המטריקה מאפשרת לבצע השוואה בין רשתות גנרטיביות שונות, או גרסאות שונות של אותו מודל. יחד עם זאת למטריקת FID קיימות מגבלות שונות. אחת מהן היא שדמיון בהתפלגות תמונות לא תמיד מתיישר עם הפרספקטיבה האנושית ואיכות ציון המטריקה לא בהכרח יעיד על כך שמודל גנרטיבי מצליח לזהות את הדפוסים בין מרחבי הדגימות.

3.5.5 בעיות בפונקציית ההפסד ברשתות גנרטיביות

תהליך הלמידה של רשת גנרטיבית כולל כאמור ייצור של דוגמאות סינטטיות על ידי המחולל אשר מתקן את עצמו באמצעות משוב המגיע מהמפלה. בעוד שהמפלה מסתמך בתהליך האימון שלו על דוגמאות אמיתיות ודוגמאות סינטטיות כדי לעדכן את פונקציית המשקל שלו, המחולל מסתמך רק על המשוב המתקבל מהמפלה. בתחילת תהליך הלמידה כאשר המחולל עדיין לא מאומן הדוגמאות הסינטטיות אותן הוא מייצר שונות מאוד מהדוגמאות האמיתיות והמפלה מפריד ביניהן בקלות. במילים אחרות בתחילת תהליך האימון המפלה טוב בהרבה מאשר המחולל. פער זה יוצר בעיה מהותית בתהליך הלימוד של המחולל כיוון ש"הידע" שעובר למחולל באמצעות הגרדיאנט של פונקציית המחיר תלוי בערכי המפלה [14].

במילים פשוטות המפלה טוב בהרבה מהמחולל בתחילת תהליך אימון ולכן האותות המתקבלים בתהליך הלימוד מהמפלה הם לא משמעותיים לתהליך הלימוד. בכדי לפתור את הבעיה הנ"ל קיימות מספר טכניקות שניתן להשתמש בהן בתהליך האימון של הרשת [14].

1. נרמול אצווה ווירטואלי – בשונה משכבת נרמול אצווה סטנדרטית (BN -Batch normalization) שבה מחושבת בכל איטרציה אימון הממוצע וסטיית התקן של כל פיצ'ר במרחב האצווה, ולאחר מכן כל פיצ'ר מנורמל באמצעותם, בשיטת נרמול באץ ווירטואלית (VBN) הנרמול מבוצע יחסית לבאץ שנשאר קבוע לכל אורך תהליך האימון. הבאץ נדגם מכל סט נתוני האימון. תהליך נרמול נתונים קבוע זה לאורך שלבי האימון עוזר לייצב את תהליך אימון ולהימנע מבעיות של קריסת מודל.
2. Minibatch discrimination – אפליה מבוססת מיני-אצווה היא שיטה נוספת שייעודה לעזור למחולל ליצור דוגמאות ריאליסטיות יותר בתהליך האימון. על פי שיטה זו, פיצ'רים נוספים מחושבים בכל אצווה אימון שמוזן למפלה. פיצ'רים אלו אמורים לכלול אינפורמציה על ההקשרים והתלויות בין דגימות שונות באותו אצוות אימון. לאחר חישוב מפת מאפיינים זו הם משורשרים אל הדוגמאות שבבאטץ הנתון ומוזרמות יחדיו אל רשת המפלה. המידע זורם ברשת המפלה ומתבצע סיווג לנתוני הקלט (נתוני אמת או נתונים סינטטיים). כך המפלה לומד לא רק על כל דוגמה בנפרד אלא גם על ההקשרים בין דוגמאות שונות ובכך יכול להיווצר מצב שבו המפלה שולח איתותים באמצעות תהליך האופטימיזציה אל המחולל המאפשר לו גם ללמוד על התלויות בין דוגמאות שונות ולהדגיש אזורים בעייתיים בתהליך יצירת התמונות שלא קיבלו "קשב" ראוי מהמחולל.

לסיכום, GANs זכו לפופולריות עצומה בעולמות עיבוד התמונה והראייה הממוחשבת. רשתות אלו נמצאות בשימוש לטובת יצירת תמונות ריאליסטיות, יצירת פרצופים ויצירות אמנות. אחד מהשימושים הנפוצים של רשתות GAN הוא בהגדלת נתונים במשימות למידת מכונה. על ידי יצירת נתונים סינתטיים, GANs עוזרים לשפר את יכולת ההכללה של מודלים, במיוחד בתרחישים שבהם השגת מערכי נתונים גדולים מסומנים היא מורכבת, לדוגמה בעולמות המידע הרפואי [8].

לצד כוחה העצום של רשת נוירונים מסוג GAN אימון רשתות GANs יכול להיות מאתגר, הן מבחינה טכנית (התמודדות עם בעיות של התכנסות האימונים או בעיות כמו היעלמות/התפוצצות שיפועים) והן מבחינה אתית. השימוש ב-GAN מעורר חששות אתיים, במיוחד ביצירת תוכן מזויף עמוק או מניפולציה של תמונות וסרטונים [8].

4. שיטות לסיווג מחלות ריאתיות באמצעות צילומי רנטגן

בפרק זה אציג ואדון בשיטות שונות לסיווג מחלות ריאתיות באמצעות תמונות רנטגן של בית החזה. טכניקות סיווג אלו מבוססות על למידה עמוקה.

הטכניקות אותן אסקור בפרק זה הן:

1. טכניקות המבוססות על סיווג באמצעות רשתות קונבולוציה בלבד.
2. טכניקות המשלבות שימוש ברשת קונבולוציה לצד רשתות נשנות.
3. טכניקות המשלבות שימוש ברשתות קונבולוציה ומנגנון הקשב.
4. טכניקות המשלבות ארכיטקטורת רשת גנרטיבית ורשת קונבולוציה.

כיוון שהמחקר בתחום התפתח בצורה מואצת בשנים האחרונות, מאגרי המידע בהם השתמשו חוקרים, שונים מאוד אחד מהשני. בין מחקרים שונים ניתן למצוא שימוש במאגרי מידע המכילים כמות שונה של תמונות לסיווג וכמות שונה של מחלקות לסיווג (קרי, מחלות שונות שתועדו בצילומי רנטגן). אי לכך אציג ואעמוד על השוני בין מאגרי המידע השונים בהם השתמשו בכל מחקר ומחקר בנפרד.

מדדי הערכה להערכת המחקרים כוללים מדדים שונים כגון accuracy, Precision, recall ומדדים נוספים. בפרק 5 ישנה סקירה מקיפה של מדדי ההערכה ובפרק 6 מבוצע ניתוח השוואתי על בסיס מדדים אלו.

בפרקים הבאים אסקור את טכניקות הסיווג השונות ומדגם מייצג של מחקרים. בכל פרק מופיע סקירה כללית של מחקרים שונים שבוצעו בהתאם לטכניקה מסוימת ובנוסף ישנה הרחבה על מקרה בוחן המייצג טכניקה מסוימת.

4.1 סיווג מחלות ריאתיות באמצעות רשתות קונבולוציה

רשתות קונבולוציה מהוות בסיס לעבודות מחקר רבות בתחום סיווג צילומי רנטגן של בית החזה לטובת אבחון מחלות ריאה.

במחקרם של [13] מוצגת ארכיטקטורה המחברת בין רשת קונבולוציה מסוג VGG19 ורשת קונבולוציה עצמאית. מטרת המחקר הייתה לבנות כלי סיווג למחלות ריאה אשר יאפשר לבצע סיווג מדויק בין 5 מחלות ריאה שונות ובין אנשים בריאים.

רשת הקונבולוציה העצמאית מורכבת משלוש שכבות קונבולוציה עם גרעינים בגדלים משתנים. לאחר שלושת שכבות הקונבולוציה הפלט מוזרם לתוך שכבת נרמול ולאחר מכן

מוזן לרשת FC בעלת ארבע שכבות סיווג בגדלים משתנים (512, 256, 128 ו-6 נירונים בהתאמה).

תהליך העיבוד המקדים פשוט וכולל שינוי גודל התמונות לגודל אחיד עבור הזנתם לרשת הנירונים ותהליך נרמול של הפיקסלים. במסגרת המחקר נבחן מאגר מידע בגודל של 47000 תמונות המחולקות לשש מחלקות סיווג שונות – חמש מחלות ריאה שונות (קורונה, דלקת ריאות, סרטן ריאות, עכירות ריאה ושחפת) ובין אנשים בריאים. מאגר המידע חולק לקבוצת אימון, אימות ומבחן כמקובל בספרות.

תוצאות המחקר מעידות כי רשת הקונבולוציה הגיעה לרמת דיוק (Accuracy) של 96.48%, (Precision) 97.56%, (Sensitivity) 93.75% על קבוצת המבחן.

במחקרם של [2] נערך שימוש ברשת קונבולוציה מסוג (50RESNET). תהליך עיבוד המידע המקדים כלל שינוי גודל תמונות הקלט לגודל אחיד של 512/512 פיקסלים (גודל גדול יחסית למחקרים אחרים בתחום) ובנוסף בוצעה התאמת קונטרסט, הפעלת פילטרים לסיווג רעשים וביצוע פעולת שוויון היסטוגרמה.

לצד תהליכי העיבוד המקדים במסגרת המחקר נבחן שימוש בראשי סיווג מסוגים שונים כגון BNN, RF, SVM לצד ראש סיווג מסוג סיגמואיד. לטובת אימות המודל בוצע שימוש בטכניקת K fold cross validation. גודל קבוצת האימון במחקר עמד על כ-4100 תמונות שחולקו בין 4 מחלקות – אנשים בריאים, חולים בקורונה, חולים בעכירות ריאה ואנשים החולים בדלקת ריאות ויראלית.

על פי תוצאות המחקר הרשת הגיעה לרמת דיוק (Accuracy) של 98.35%, (Precision) 97.43%, (Sensitivity) 98.87% על קבוצת המבחן.

גם מחקרם של [11] מיושם מנגנון עיבוד תמונה מקדים לתמונות הקלט טרם השימוש ברשת קונבולוציה מסוג 19VGG ורשתות קונבולוציה שנבנו באופן עצמאי. מנגנון עיבוד התמונה המקדים שמוצג במחקר זה מכונה CLAHE. מנגנון זה מבוסס על עקרון שוויון היסטוגרמה. על פי מנגנון זה, יש לבצע שוויון היסטוגרמה על קטעי תמונה שאינם חופפים, ולאחר מכן לבצע אינטרפולציה בכדי לתקן אנומליות בין הגבולות של המקטעים. לצד שיטת העיבוד המקדים והשימוש ברשתות קונבולוציה, בחנו החוקרים שימוש במנגנוני סיווג מתחום הראייה הממוחשבת שאינם בנויים מרשתות נירונים כגון אלגוריתם SVM.

תוצאות המחקר הראו הבדלי דיוק משמעותיים בין רשתות נירונים (91% דיוק על קבוצת המבחן) למסווג SVM (67.2%).

במסגרת המחקר בוצע שימוש בדאטה סט בגודל של כ-21000 תמונות שחולקו לארבע מחלקות – אנשים החולים בקורונה, אנשים החולים בדלקת ריאות, אנשים בריאים ואנשים החולים בעכירות ריאות.

בין המחקרים השונים ניתן למצוא בסיס משותף. כלל המחקרים עוסקים בבחינת היעילות של רשתות המבוססות על קונבולוציה ללא שימוש בשכבות נוספות כגון שכבות נשנות או מנגנון קשב. השוני בין המחקרים נובע משיטות עיבוד התמונה טרום הזנתן לרשת הניורונים או בסוג ועומק רשתות הקונבולוציה שנבחנו. בעבודה זו בחרתי להתמקד ולהדגים את שיטת הסיווג המבוססת על קונבולוציה בלבד באמצעות מקרה בוחן מייצג.

4.1.1 מקרה בוחן לסיווג מחלות ריאה באמצעות רשתות קונבולוציה

בעבודתם של [5] מוצגת שיטה לסיווג מחלות ריאה על בסיס שימוש ברשת קונבולוציה. החוקרים התבססו על כך ששימוש ברשתות קונבולוציה לניתוח צילומי רנטגן של בית החזה הוכיחו את עצמן כיעילות ובעלות פוטנציאל יישומי בניתוח מחלות ריאתיות אחרות שנחקרו בעבר.

מטרתם של החוקרים הייתה לסווג בין תמונות של אנשים החולים בקורונה ובין אנשים בריאים ללא התייחסות למחלות ריאתיות אחרות תו"כ חקירת רשתות קונבולוציה שונות. החוקרים פיתחו מספר רשתות קונבולוציה בגדלים משתנים ועקבו אחר השינויים ברמות הדיוק שרשתות אלו מציגות, לאחר מכן השוו את תוצאותיהם לתוצאות רשתות ניורונים מוכרות מסוג 19VGG.

מלבד השוואת הארכיטקטורה של רשתות קונבולוציה בגדלים שונים, החוקרים בחנו מספר ראשי סיווג שונים. רמת הדיוק של המודלים השונים היוותה את כלי ההערכה המרכזי במחקר.

4.1.2 מאגר הנתונים

גודל מאגר הנתונים במחקר זה היה יחסית מוגבל וכלל 380 נבדקים. הנתונים תוייגו ע"י צוות רפואי. מאגר המידע הכיל 180 צילומי רנטגן של אנשים החולים בקורונה וב-200 תמונות של אנשים בריאים. גודל כל תמונה קבוע ועומד על 224/224/3 פיקסלים. מרחב הדגימה במחקר מצומצם מאוד יחסית לבעיות דומות בעולם הלמידה העמוקה.

4.1.3 שיטת המחקר

המחקר התחלק כאמור לשלוש חלקים שונים. בשלב הראשון פיתחו החוקרים מספר רשת נוירונים קונבולוציוניות ייחודיות לטובת סיווג התמונות. תהליך המחקר כלל פיתוח ובחינה של מספר תתי רשתות שונות בעלות עומקים שונים. כל תת רשת אומנה ובחנה על מאגר המידע בצורה נפרדת. מבנה הרשתות הלך והעמיק עם התקדמות המחקר וכך גם רמת הביצועים של הרשתות. באיור מטה ניתן לראות את ההבדל בין סוגי הרשתות שנחקרו ואת רמת הדיוק עבור כל אחת מהרשתות:

CNN Model	- input - conv_1 - BN_1 - relu_1 - pool_1 - conv_2 - BN_2 - relu_2 - pool_2 - conv_3 - BN_3 - relu_3 - conv_4 - BN_4 - relu_4 - conv_5 - BN_5 - relu_5 - FullyConnected - Softmax - Classification_layer	- input - conv_1 - BN_1 - relu_1 - pool_1 - conv_2 - BN_2 - relu_2 - conv_3 - BN_3 - relu_3 - conv_4 - BN_4 - relu_4 - FullyConnected - Softmax - Classification_layer	- input - conv_1 - BN_1 - relu_1 - pool_1 - conv_2 - BN_2 - relu_2 - pool_2 - conv_3 - BN_3 - relu_3 - FullyConnected - Softmax - Classification_layer
Accuracy	91.58	88.42	86.32

איור 25 -רשתות קונבולוציה [5].

ניתן לראות שהרשתות הורכבו ממספר משתנה של שכבות, אך המבנה הבסיסי וסוג השכבות השונות נשמר בין מקרי הבחינה השונים. בעוד שהרשת הראשונה הורכבה מ15 שכבות שונות והגיעה לרמת דיוק של 86.32%, הרשת הסופית שנבחנה נבנתה מ21 שכבות והגיעה לרמת דיוק של 91.58%. הרשתות נבנו כולן מאותן "אבני בניין" – שכבות המבצעות קונבולוציה לצד שכבות איגום ושכבות נרמול (batch normlization - BN). בסוף כל אחת מהרשתות נמצאת שכבת FC ושכבת softmax ולאחריה שכבת קלסיפיקציה.

הרשת הסופית שהוצעה הורכבה מ21 שכבות. הרשת כללה כאמור שכבות קונבולוציה בגדלים שונים, שכבות איגום ושכבות נרמול. גודל ופירוט על השכבות כמפורט באיור מטה:

1	"input"	$224 \times 224 \times 3$ images
2	"conv_1"	$64 \ 3 \times 3 \times 3$ convolutions with stride [1 1] and padding "same"
3	"BN_1"	Batch normalization
4	"relu_1"	ReLU
5	"pool_1"	2×2 max pooling with stride [2 2] and padding [0 0 0 0]
6	"conv_2"	$32 \ 3 \times 3 \times 8$ convolutions with stride [1 1] and padding "same"
7	"BN_2"	Batch normalization
8	"relu_2"	ReLU
9	"pool_2"	2×2 max pooling with stride [2 2] and padding [0 0 0 0]
10	"conv_3"	$16 \ 3 \times 3 \times 16$ convolutions with stride [1 1] and padding "same"
11	"BN_3"	Batch normalization
12	"relu_3"	ReLU
13	"conv_4"	Eight $3 \times 3 \times 16$ convolutions with stride [1 1] and padding "same"
14	"BN_4"	Batch normalization
15	"relu_4"	ReLU
16	"conv_5"	Four $3 \times 3 \times 16$ convolutions with stride [1 1] and padding "same"
17	"BN_5"	Batch normalization
18	"relu_5"	ReLU
19	"fc"	Five fully-connected layers
20	"softmax"	Softmax
21	"Classification"	crossentropyx with "0" and nine other classes

איור 26 - רשת ניורונים [5].

הרשת מקבלת כקלט תמונה בגודל של $224/224/3$ פיקסלים. המידע זורם לאחר מכן לשכבת קונבולוציה בעלת 64 גרעינים בגודל של $3/3$ עם גודל צעד של 1. לאחר הקונבולוציה המידע זורם לשכבת נרמול ולפונקציית אקטיבציה מסוג Relu. בסוף בלוק הקונבולוציה הראשון נמצאת שכבת איגום.

התהליך חוזר על עצמו לאורך הרשת כאשר כמות הגרעינים משתנה בין שכבות הקונבולוציה השונות אך שכבות הנרמול, האקטיבציה ואיגום נשארות זהות. בסוף פעולת הקונבולוציה לאורך הרשת המידע מוזרם לשכבת FC ולשכבת softmax לטובת סיווג התמונות.

החוקרים השתמשו באלגוריתם אופטימיזציה מסוג מורד גרדיאנט סטוכסטי (SGD). הרשת אומנה לאורך 300 אפוקים.

בחלק השני של המחקר בחנו החוקרים את ההשפעות של ראשי סיווג שונים בכדי למקסם את יכולות הסיווג של רשת הניורונים. החוקרים בדקו את ההשפעות של מסווג SVM מסוגים שונים (linear, quadratic, cubic, gaussian).

בחלק השלישי השוו את תוצאות הרשת הייחודית שבנו לרשתות קונבולוציה אחרות באמצעות שימוש בלמידת העברה. הרשתות שנבחנו היו RESNET19, RESNET50, RESNET101, VGG16, VGG19.

החוקרים הצליבו את ראשי הסיווג השונים שנחקרו עם רשתות הקונבולוציה שהשתמשו בהם במטרה למקסם את איכות הדיוק של הרשתות. כל ראש סיווג נבחן לצד כל אחת מרשתות הקונבולוציה.

Method: SVM	Accuracy (%)					
	ResNet18	ResNet50	ResNet101	VGG16	VGG19	Average
Linear Kernel	86.3	94.7	88.4	89.5	88.4	89.5
Quadratic Kernel	87.4	91.6	89.5	89.5	87.4	89.1
Cubic Kernel	89.5	90.5	91.6	90.5	89.5	90.3
Gaussian Kernel	86.3	93.7	88.4	89.5	87.4	89.1
Average	87.4	92.6	89.5	89.8	88.1	

טבלה 1 - תוצאות ראשי הסיווג עם רשתות הקונבולוציה [5].

באיוור ניתן להבחין בתוצאות ההצלבה של ראשי הסיווג שונים עם רשתות הקונבולוציה השונות. מימד השורות בטבלה מציין את סוג ראש הסיווג מסוג SVM, ומימד העמודות את סוג רשת הקונבולוציה. התוצאות המספריות מייצגות את רמת הדיוק שהגיעו אליה החוקרים בהצלבת כל אחת מהשיטות.

4.1.4 תוצאות המחקר ומסקנות

על פי המחקר התוצאות הטובות ביותר הגיעו מרשת הקונבולוציה 50RESNET בשילוב עם ראש סיווג מסוג SVM לינארי. רמת הדיוק שהוצגה היא של 94.7%. אימון של רשת 50RESNET בלבד, ללא ראש סיווג SVM הביא לתוצאות של 92.6%, ורשת הקונבולוציה הייחודית הביאה לתוצאות של 91.6%. התוצאות של המחקר מוצגות באיוור מטה.

Method	Accuracy (%)
ResNet50 Features + SVM	94.7
Fine-tuning of ResNet50	92.6
End-to-end training of CNN	91.6

טבלה 2 - תוצאות המחקר [5].

4.2 סיווג מחלות ריאתיות באמצעות רשתות CNN-RNN

רשתות קונבולוציה המשולבות עם רשתות נשנות הן רשתות נוירונים מורכבות אשר משלבות בין סוגים שונים של רשתות נוירונים עמוקות. רשתות אלו מורכבות מתת רשת המבצעת את פעולות הקונבולוציה, ומתת רשת נשנת נוספת. מבנה זה של רשת הנוירונים השלמה מנסה לשפר את תהליך הוצאת המאפיינים מתמונת הקלט.

מחקרם של [18] מתמקד בבחינת השילוב בין רשתות קונבולוציה עמוקות ובפרט הרשתות Vgg 19, Densnet 121, Resnet 152 לצד שילובן עם שכבות נשנות מסוג GRU, LSTM (GRU) הם סוג של תא רשת נשנת המבוסס על ארכיטקטורת LSTM עם מספר שינויים פנימיים האמורים לייעל את תהליך העיבוד הפנימי בתוך התא).

המחקר בוחן את השפעת החיבור של רשתות קונבולוציה עם שכבות נשנות. במחקר מוצג תהליך עיבוד תמונה הכולל שימוש באלגוריתם CLAHE שמבוסס על שוויון היסטוגרמה, לצד שימוש בפילטר מסוג מסנן גאמא – פילטר אשר מתמודד עם זוהר התמונה באמצעות טרנספורמציה לא ליניארית.

המחקר התבסס על מאגר מידע בגודל של כ-16000 תמונות רנטגן CTI של בית החזה המחולק ל-3 מחלקות – אנשים בריאים, חולים בקורונה וחולים בדלקת ריאות. המודל המשלב בין רשת GRU + resnet152 הביאה לרמת הביצועיים הטובה ביותר, 93.37% דיוק (Accuracy), 93.73% (precision), ו-93.47% (sensitivity).

במחקרם של [10] נחקר שילוב של רשתות קונבולוציה טהורות מסוג VGG19 אשר חוברו עם שכבות קונבולוציה נוספות לצד השוואתם לרשתות נוירונים המשלבות רשתות קונבולוציה מסוג resnet152V עם שכבות נשנות מסוג GRU-BI (דו כיווני), GRU-BI (דו כיווני) GRU-BI היא שכבה נשנת מסוג GRU עם שינוי ייחודי – הקלט מוזרם לשכבה זו בשתי דרכים שונות.

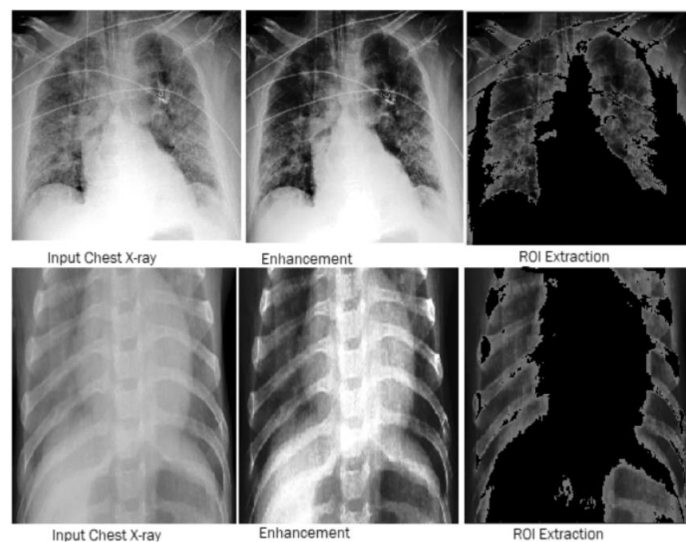
ראשית מוזרם הקלט אל השכבה כרגיל, קרי מראשיתו ועד סופו, ושנית מוזרם הקלט במהופך קרי מסופו ועד תחילתו. בצורה זו מנסה שכבת נוירונים ייחודית זו למצוא דפוסים ייחודיים נוספים בנתוני הקלט הן בתחילתם והן בסופן.

גודל קבוצת המדגם בו השתמשו במחקר בעל 33676 תמונות המחולקות לארבע מחלקות סיווג – אנשים בריאים, אנשים החולים בקורונה, אנשים החולים בדלקת ריאות ואנשים החולים בסרטן ריאה.

תוצאות המחקר מראות כי רשתות הקונבולוציה הטהורות המבוססות על רשת VGG19 בתוספת שכבות קונבולוציה נוספות עקפו את רשתות הנוירונים המשלבות שכבות נשנות.

התוצאות המדווחות במחקר מעידות על דיוק של 98.5% (Accuracy), 99.5% (specificity), ו-98.43% (precision).

מחקר נוסף של [25] מציג שיטת עיבוד תמונה ייחודית טרום הזנת הקלט לרשת הנירונים, המכונה ROI extraction. הרעיון העומד מאחורי שיטת עיבוד תמונה זו היא ההנחה שהמידע המשמעותי ביותר בתמונת הקלט עבור סיווג מחלות ריאה נמצא בריאות עצמן ולא בבית החזה כולו. באמצעות תהליך של סגמנטציה המזהה את גבולות הריאה בתוך בית החזה, תהליך העיבוד המקדים מתמקד בריאות ומוחק מהתמונה אזורים שאינם אינפורמטיביים לסיווג התמונות. כך המידע הזורם אל רשת הנירונים הוא ממוקד יותר ונקי מרעשים. בנוסף במחקר זה מבוצע עיבוד מקדים נוסף באמצעות פילטר חציוני ושוויון היסטוגרמה.



איור 27 - שימוש באלגוריתם ROI לזיהוי הריאות [25].

מלבד שיטת העיבוד המקדים של תמונות הקלט, המחקר בוחן את הדיוק בסיווג התמונות בין טכניקות שונות של למידת מכונה כמון אלגוריתם KNN, SVM לצד שימוש ברשתות קונבולוציה המשולבות עם שכבת LSTM. מאגר המידע מחולק לארבע קטגוריות – אנשים החולים בדלקת ריאות וויראלית, אנשים החולים בדלקת ריאות בקטריאלית, אנשים בריאים ואנשים החולים בקורונה. מאגר המידע מכיל כ-8500 תמונות.

תוצאות המחקר מראות כי רמת הדיוק הטובה ביותר נצפתה עם מסווג המבוסס על רשת קונבולוציה משולבת עם שכבת LSTM, ואחוז הדיוק אליו הגיעו במחקר עומד על 94.31% (Accuracy).

בחרתי להתמקד במקרה בוחן המתואר במחקר של [31] הבוחן שימוש ברשת קונבולוציה לצד שכבת LSTM.

4.2.1 מקרה בוחן לסיווג באמצעות מודל CNN-LSTM

[31] מציעים דרך שונה להתמודדות עם סיווג צילומי רנטגן עבור חולים במחלות ריאה. במחקרם הם משלבים רשת קונבולוציה עם שכבות LSTM. מחקרם מבוסס על מדגם נתונים רחב, אשר באמצעותו מסווגות תמונות הקלט בין מספר מחלות ריאה שונות. במחקר מבוצעת השוואה בין רשת CNN-LSTM לרשת קונבולוציה טהורה. מטרת המחקר הייתה לבנות כלי לסיווג תמונות רנטגן של הריאות בין אנשים החולים בקורונה, דלקת ריאות ובין אנשים בריאים.

4.2.2 מאגר הנתונים

מאגר הנתונים הורכב מ-4575 צילומי רנטגן. גודל כל תמונה הוא 224/224/3 פיקסלים. באיור מטה ניתן לראות את חלוקת התמונות לפי הסיווג שלהן במחקר, מחולקים לפי קבוצת מבחן וקבוצת אימון.

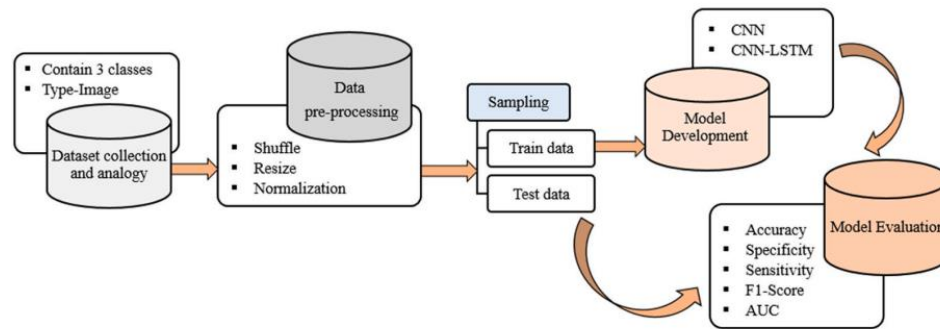
Data/Cases	COVID-19	Normal	Pneumonia	Overall
Training	1220	1220	1220	3660
Testing	305	305	305	915
Overall	1525	1525	1525	4575

טבלה 3 - תיאור הנתונים [31].

4.2.3 שיטת המחקר

שיטת המחקר כללה תהליך טיוב ושיפור של מודל CNN-LSTM תחילה טיבו החוקרים את תהליך האוגמנטציה והנרמול. כל תמונה עברה תהליך של נרמול ואוגמנטציה, בנוסף, בשל העובדה שהתמונות נאספו ממאגרי מידע שונים, כל אחת מהתמונות עברה תהליך של שינוי גודל התמונה בכדי להתאים את הנתונים לגודל אחיד. לאחר תהליך ההכנה של הנתונים, הנתונים חולקו לקבוצת אימון וקבוצת מבחן.

הנתונים הוזרמו לשתי רשתות נוירונים שונות, שאומנו בצורה נפרדת. כאמור רשת נוירונים אחת הורכבה משכבות קונבולוציה בלבד, ולרשת השנייה נוספה שכבת LSTM בסוף שכבות הקונבולוציה. לאחר אימון המודלים השונים, בוצעה הערכה לכל אחד מהמודלים בנפרד.



איור 28 - תיאור סכמטי של תהליך אימון המודלים [31].

נתמקד ברשת CNN-LSTM ברמה התאורטית, ברשת זו תפקידן של שכבות הקונבולוציה הוא לשמש ככלי להוצאת פיצ'רים ולבצע עיבוד של תמונת הקלט. ותפקידה של שכבת הLSTM הוא לשמש לשפר את תהליך הוצאת הפיצ'רים על ידי שימוש בבחינת הקלט לאורך רצף נתונים משתנה.

תמונה – המיוצגת על ידי שלושה מימדים (מימד הרוחב, הרוחב והעומק) היא איננה מתאימה להזנה לשכבת LSTM המצפה לקבל רצף נתונים משתנה לאורך מימדים קבועים. בכדי להזין את תוצאות רשת הקונבולוציה לשכבת LSTM נדרש לשנות את מימדי הפלט של רשת הקונבולוציה לכדי שתתאים לרשת LSTM.

הקלט לשכבת הקונבולוציה הן תמונות בגודל 224/224/3 פיקסלים. פעולות רשת הקונבולוציה מצמצמת את מימד השורות ומימד העמודות של התמונה מחד אך מגדילה את כמות הערוצים של התמונה. במילים אחרות, לצד דחיסה בגודל התמונה (שמיצג על ידי מימד השורות והעמודות) גדל מימד הערוצים של התמונה (המיצג על ידי המימד השלישי שלה). בסוף תהליך הקונבולוציה נוצרת שכבת פלט בגודל של 7/7/512 פיקסלים.

במילים אחרות פעולת הקונבולוציה על תמונת הקלט מייצרת רצף גדול מספיק של תמונות פלט בגודל קטן יחסית (7/7 פיקסלים לכל תמונת פלט). תמונות הפלט של רשת הקונבולוציה אמורות לייצג את תמונת הקלט.

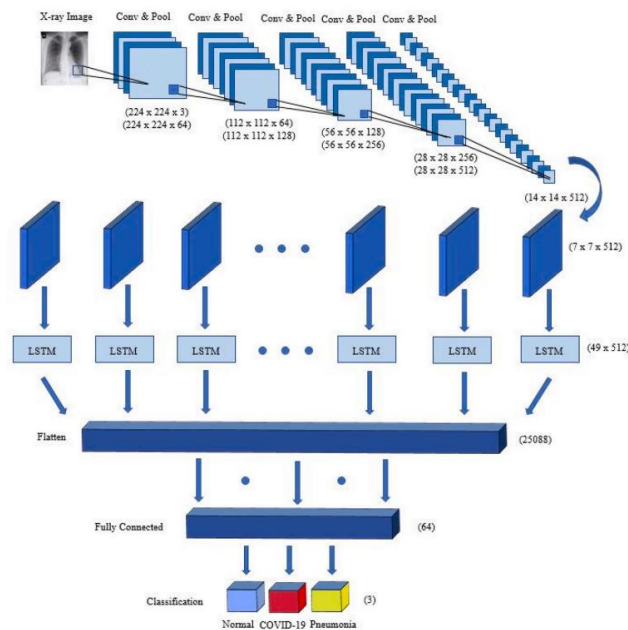
ברשתות שאינן מחוברות לשכבת סיווג מסוג LSTM, פלט שכבת הקונבולוציה היה נפרש למימד אחד ארוך, אשר בתורו היה מועבר לשכבת FC לטובת סיווג. אך לא כך הדבר כאשר משתמשים בשכבת LSTM.

נשאלת השאלה – כיצד פלט שכבת הקונבולוציה יכול להיות בשימוש של שכבת LSTM – שהרי פלט שכבת הקונבולוציה הוא אוסף רחב של תמונות ממוזערות, וקלט שכבת הLSTM הוא רצף של נתונים המשתנים לאורך זמן.

את רצף התמונות המתקבלות כפלט מרשת הקונבולוציה ניתן להזין לשכבת ה-LSTM בצורה הבאה – כל אחת מתת התמונות הקטנות הן חלק מרצף הנתונים המזין את שכבת ה-LSTM. רצף זה אומנם לא משתנה לאורך הזמן, אך הוא משתנה כתלות בגרעינים שהופעלו עליו לאורך שכבות הקונבולוציה. כלומר ישנו רצף של תמונות המשתנות לאורך מימד הפיצ'רים של רצף התמונות.

כל אחד מהפיקסלים בכל אחת מתת התמונות מהווה פיצ'ר שמוזרם לתוך שכבת ה-LSTM. בכך מקבלת שכבת ה-LSTM רצף של תמונות אשר ניתן לעיבוד וניתוח. הרשת למעשה בוחנת את השינוי והקשר שנוצר בין רצף תמונות הפלט של שכבות הקונבולוציה על ידי שימוש בשכבת ה-LSTM.

בסוף תהליך העיבוד של שכבת ה-LSTM מחברים ראש סיווג נוסף המורכב משכבת FC כמקובל בתחום.



איור 29 - ארכיטקטורת רשת CNN-LSTM [31].

The full summary of CNN-LSTM network.

Layer	Type	Kernel Size	Stride	Kernel	Input Size
1	Convolution2D	3×3	1	64	$224 \times 224 \times 3$
2	Convolution2D	3×3	1	64	$224 \times 224 \times 64$
3	Pool	2×2	2	–	$224 \times 224 \times 64$
4	Convolution2D	3×3	1	128	$112 \times 112 \times 64$
5	Convolution2D	3×3	1	128	$112 \times 112 \times 128$
6	Pool	2×2	2	–	$112 \times 112 \times 128$
7	Convolution2D	3×3	1	256	$56 \times 56 \times 128$
8	Convolution2D	3×3	1	256	$56 \times 56 \times 256$
9	Pool	2×2	2	–	$56 \times 56 \times 256$
10	Convolution2D	3×3	1	512	$28 \times 28 \times 256$
11	Convolution2D	3×3	1	512	$28 \times 28 \times 512$
12	Convolution2D	3×3	1	512	$28 \times 28 \times 512$
13	Pool	2×2	2	–	$28 \times 28 \times 512$
14	Convolution2D	3×3	1	512	$14 \times 14 \times 512$
15	Convolution2D	3×3	1	512	$14 \times 14 \times 512$
16	Convolution2D	3×3	1	512	$14 \times 14 \times 512$
17	Pool	2×2	2	–	$14 \times 14 \times 512$
18	LSTM	–	–	–	49×512
19	FC	–	–	64	25,088
20	Output	–	–	3	64

איור 30 - המודל המלא CNN-LSTM - [31].

באיור מעלה ניתן לראות את ארכיטקטורת רשת הניורונים המוצעת במחקר. הרשת בנויה משכבות קונבולוציה בעלות קרנל בגודל של $3/3$ עם מספר ערוצים הולך וגדל. בין כל זוג שכבות קונבולוציה מבוצעת פעולת איגום. לאחר מעבר בשלוש "בלוקים" שכאלו, משתנה במעט מבנה הרשת, ונוספת שכבת קונבולוציה נוספת לפני ביצוע פעולת האיגום. לאחר שני "בלוקים" שכאלו מוזרם הפלט של רשת הקונבולוציה לשכבת LSTM ולאחר מכן מועבר לשכבת FC לצורך סיווג.

4.2.4 תוצאות המחקר ומסקנות

על פי תוצאות המחקר, הרשת המשולבת CNN-LSTM הגיע לאחוזי דיוק מרשימים. יש לזכור שעל אף התוצאות המרשימות גם במחקר זה גודל הנתונים היה יחסית קטן עבור מודל של למידה עמוקה. ניתן להניח שבחינת הרשת עם מאגר מידע גדול יותר ובעל מחלקות סיווג נוספות יביא כנראה לירידה באחוז הדיוק של הרשת המוצעת וזאת בשל העובדה שכולל שגודל המדגם גדול יותר, כך הוא מכיל דפוסים רבים ושונים יותר. לרשת הניורונים אם כן יהיה קושי רב יותר להכיל את כלל הדפוסים והייצוגים של המידע. אי לכך ניתן לצפות כי שימוש במאגר מידע גדול יותר, ייפגע בביצועי הרשת.

Performance of the CNN-LSTM network.

Class	Accuracy (%)	Specificity (%)	Sensitivity (%)	F1-score (%)
COVID-19	99.2	99.2	99.3	98.9
Pneumonia	99.2	99.8	98.0	98.8
Normal	99.8	99.7	100.0	99.7

טבלה 4 - תוצאות המחקר [31].

באזור מעלה ניתן לראות את תוצאות המחקר. מימד השורות מציין את סוג המחלקה המסווגת, ומימד העמודות את סוג מדד הערכה שנמדד.

4.3 סיווג מחלות ריאתיות באמצעות רשת קונבולוציה ומנגנון קשב

לצד המחקר הבוחן שילוב רשתות קונבולוציה עם רשתות נשנות, מבוצע מחקר נוסף הבוחן את יעילותו של שילוב רשתות קונבולוציה עם מנגנון הקשב.

[26] מציגים שילוב מורכב יחסית בין רשתות קונבולוציה ומנגנון הקשב. במחקרם עיצבו החוקרים רשת נוירונים הבנויה משלוש תת רשתות המבצעות את פעולת הקונבולוציה (Resnet – 152, Resnet – 18, Densnet).

תמונת הקלט זורמת בנפרד בתוך שלושת תתי רשתות הקונבולוציה במקביל ולאחר שלב עיבוד מידע זה הקלט משורשר יחדיו לווקטור אחד שממדו 3/1/1024. ווקטור זה מוזן לאחר מכן לתוך שכבת קשב וממשיך לתוך ראש סיווג מסוג FC. במהלך המחקר בוצע שימוש בשיטות אוגמנטציה רבות כגון סיבוב וחיתוך התמונה במקומות שונים. במחקר זה בוצע סיווג בין אנשים בריאים ובין אנשים החולים בדלקת ריאות בלבד. גודל מאגר המידע עמד על 6000 תמונות. המחקר מציג תוצאות של 96.63% (Accuracy), 96.24% (Precision), ו-98.64% (recall).

ניתן למצוא בתת תחום זה גם מחקרים המשלבים בין שכבות קשב, קונבולוציה ושכבות חוזרות כגון מחקרו של [30] במחקר זה מבוצע שילוב בין מנגנון הקשב לרשת נוירונים המורכבת משכבות קונבולוציה ושכבות נשנות. לאחר תהליך העיבוד המקדים של תמונות הקלט, המידע מועתק ומפוצל לשני חלקים בדומה לצורה של בלוק שיורי. חלקו הראשון של המידע זורם דרך רשת קונבולוציה ומנגנון קשב. וחלקו השני של המידע יתאחד עם המידע שעבר עיבוד לאחר פעולת שכבת הקשב.

רשת הקונבולוציה מורכבת משלושה בלוקים, כאשר כל בלוק מורכב משכבת קונבולוציה, שכבת נרמול, פונקציית אקטיבציה מסוג Relu, ושכבת איגום. בסוף רשת הקונבולוציה, הקלט מפוצל ומועתק פעם נוספת. חלקו הראשון של פלט יוזרם לתוך שכבת הקשב, וחלקו השני יאוחד עם פלט שכבת הקשב ולא יעבור עיבוד בתוכה. בסוף שכבת הקשב מבוצע אם כן שרשור של שלושת הפלטים – הפלט הראשון שלא עבר תהליך עיבוד כלל לא ברשת הקונבולוציה ולא בשכבת הקשב, ביחד עם הפלט שעבר עיבוד רק באמצעות רשת הקונבולוציה ובנוסף עם הקלט שעבר את תהליך העיבוד השלם.

הפלט המשורשר מוזן לאחר מכן לשכבת LSTM ולאחר מכן מועבר לתוך שכבת FC לצורך ביצוע הסיווג.

מנגנון מורכב זה מאומן באמצעות מאגר המידע במחקר עמד על 1061 תמונות המחולקות לשלוש מחלקות – אנשים בריאים, אנשים החולים בקורונה ואנשים החולים בדלקת ריאות.

על פי טענת החוקרים, רשת נוירונים זו מגיעה לדיוק של 100% (Accuracy) על קבוצת המבחן. יש לזכור שגודל מאגר הנתונים הינו קטן יחסית ושכמות המחלקות המסווגות במחקר מצומצם אף הוא. מנקודת מבטי תוצאה זו אינה נראית הגיונית ויכול להיות שקיים כשל במחקר. הכשל יכול לנבוע מגודל קבוצת המדגם או מבחירת התמונות שהרכיבו את קבוצת המבחן. על פי רוב, מחקרים בתחום אינם מצליחים לרמות דיוק מושלמות אלא רק לאחוזי דיוק גבוהים במיוחד. הגעה לדיוק מושלם על קבוצת המבחן יכולה לרמוז שהמודל שנבנה במהלך המחקר פותר את בעיית סיווג התמונות של חולים במחלות ריאה שונות בצורה מושלמת.

בפרק זה בחרתי להתמקד במחקרם של [12] אשר בוחן את השימוש בין חיבור רשת קונבולוציה ומנגנון הקשב.

4.3.1 מקרה בוחן לסיווג עם רשת קונבולוציה המשולבת מנגנון הקשב

במאמרם של [12] מוצגת רשת נוירונים מורכבת לסיווג תמונות. הרשת משלבת בין רשת נוירונים המבצעת קונבולוציה ובין שכבות קשב. המידע ברשת זורם במספר ערוצים עד להגעתו לראש סיווג.

במחקרם בחנו החוקרים את הסיווג של צילומי רנטגן של בית החזה לצד סיווג תמונות CT של בית החזה. על אף שבמחקר הוצגו תוצאות המשלבות את עבודת הסיווג על גבי סריקות רנטגן וCT במקביל, אתייחס בסקירה זו רק לחלק בעבודתם המתמקד בסריקות הרנטגן וזאת בכדי לנסות ולבודד את סוג הנתונים בו השתמשו החוקרים ובכדי לאפשר סקירה השוואתית בין השיטות השונות שנבחנו בעבודה זו.

במחקר זה בוצע ניסיון לשל את מנגנון הקשב במקביל לשכבות קונבולוציה, בכדי להביא לשיפור ביצועים בהקשר זמן האימון של המודל, והיכולת לפרוס את המודל במקומות בהם כוח העיבוד החומרתי הוא דל יחסית. השיפור בביצועים אינו יכול לפגוע ביעילות רשת הנוירונים.

לצד הניסיון להביא לאחוזי דיוק גבוהים ניסו החוקרים להתמקד ברשתות נוירונים קטנות יחסית, בכדי לנסות ולמזער את כוח העיבוד והחישוב הדרוש לביצוע משימת סיווג התמונות וזיהוי מחלות הריאה. מלבד הצמצום בכוח החישוב ניסו החוקרים להקטין את משך הזמן הדרוש לביצוע אימון של מודל שלם ושל סיווג תמונה בודדת. על ידי כך מנסים החוקרים לתת מענה למרפאות קטנות שלא בטוח שמציידות בכוח חישובי מספק לשימוש במודלים כבדי משקל.

4.3.2 מאגר הנתונים

המחקר כלל 17439 תמונות RBG של צילומי רנטגן של בית החזה. במחקר בוצע סיווג בין ארבע מחלקות – אנשים בריאים, אנשים החולים דלקת ריאות בקטריאלית, אנשים החולים בדלקת ריאות חיידקית ואנשים החולים בקורונה.

המידע הגולמי במאגר הנתונים היה לא מאוזן (כמות התמונות לכל אחת ממחלקות הסיווג הייתה לא שווה) דבר שיכול לעודד את התופעה שבה המודל ייטה לכיוון המחלקות הנמצאות בייצוג ייתר בנתונים סט הגולמי. אי לכך נעשה שימוש במחקר בכלי אוגמנטציה שכללו שימוש בהתמרות אפניות וסיבוב התמונה על צירה האופקי.

באיור מטה ניתן לראות את חלוקת הנתונים לפי מחלקות לפני ביצוע האוגמנטציה ואחריה.

Group	Category	Before	After	Category	Before	After
Four classes (total: 17,439)	Normal_X	2916	4832	Virus_X	1733	4166
	Bacteria_X	2705	4418	Covid-19_X	1341	4023

טבלה 5 - חלוקת הנתונים סט לפי מחלקות [12]

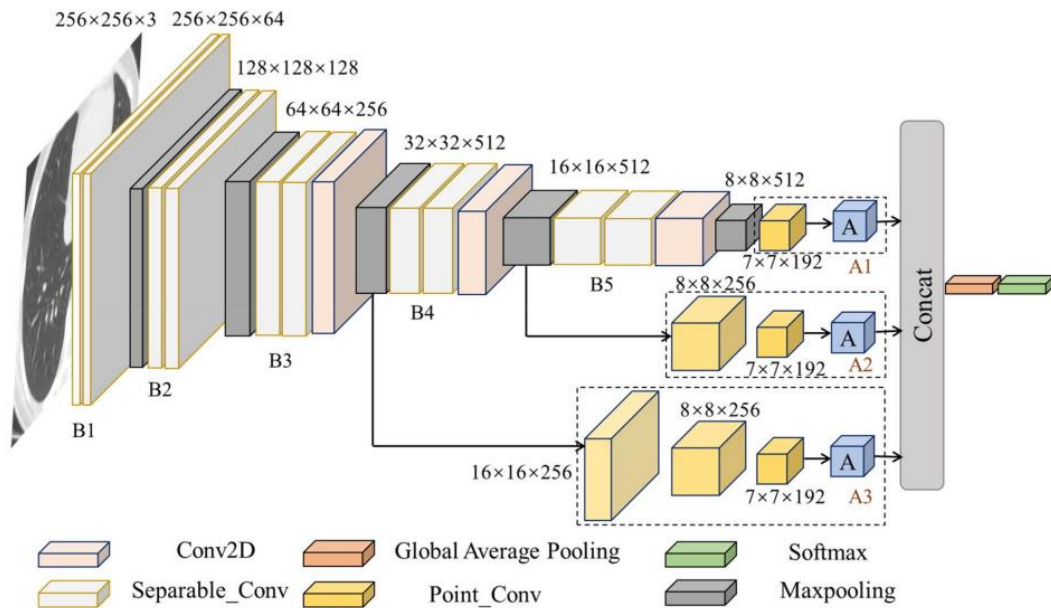
4.3.3 מהלך המחקר

רשת הניורונים שהוצע במהלך המחקר היא שילוב של רשת קונבולוציה מורכבת עם שכבות של מנגנון קשב אשר קיבלה את השם 3-MGMADS.

שני הבלוקים הקונבולוציוניים הראשונים (B1 ו-B2 בטבלה מטה) מורכבים משתי שכבות קונבולוציה ושכבת איגום. בעקבותיהם באים שלושה בלוקים קונבולוציוניים (B3, B4 ו-B5 בטבלה) הכוללים שלוש שכבות קונבולוציה מסוגים שונים ושכבת איגום.

בסרטוט ניתן לראות שבסוף בלוק B3 ובלוק B4 הנתונים מוזרמים דרך שני חלקים שונים ומופרדים של רשת הניורונים. מחד, בסוף הבלוק B3 הנתונים מוזרמים לבלוק B4, מאידך הנתונים מוזרמים במקביל ובצורה נפרדת לגמרי לבלוק נוסף העונה לשם בלוק 3A. התהליך קורה גם בסוף בלוק B4 כך שבסוף הבלוק הנתונים מוזרמים מחד לבלוק B5 וגם לבלוק A2. הבלוק B5 מזרים את המידע בסופו לבלוק A3.

הבלוקים A1-3 הם בלוקים המשלבים שכבות קונבולוציה שבסופם שכבת קשב. כלל הנתונים המוזרמים ברשת יוזרמו לתוך שכבות הקשב השונות בהתאם למקום בו בוצע הפיצול בזרימת המידע. לאחר הזרמת המידע בשכבות הקשב, מתבצע שרשור של כלל המידע שעובד בבלוקים השונים ומתבצע פרדיקציה וסיווג באמצעות ראש סיווג מסוג FC.



איור 31 - תיאור רשת הניורונים [12].

Blks	Layer (type)	Output shape	Paras	Blks	Layer (type)	Output shape	Paras
	Input layer	(None, 256, 256, 3)	0		Separable_Conv2D_16	(None, 16, 16, 512)	133,888
B1	Separable_Conv2D_1	(None, 256, 256, 64)	283	B5	Separable_Conv2D_17	(None, 16, 16, 512)	267,264
	Separable_Conv2D_2	(None, 256, 256, 64)	4736		Conv2D_18	(None, 16, 16, 512)	2,359,808
	Maxpooling2d_1	(None, 128, 128, 64)	0		Maxpooling2d_5	(None, 8, 8, 512)	0
					Point_Conv2D_19	(None, 7, 7, 192)	393,408
B2	Separable_Conv2D_3	(None, 128, 128, 128)	8896	A1	Multi-head attention_A_1	(None, 49, 32)	117,344
	Separable_Conv2D_4	(None, 128, 128, 128)	17,664		Point_Conv2D_14	(None, 8, 8, 256)	131,328
	Maxpooling2d_2	(None, 64, 64, 128)	0		Point_Conv2D_15	(None, 7, 7, 192)	196,800
B3	Separable_Conv2D_5	(None, 64, 64, 256)	34,176	A2	Multi-head attention_A_2	(None, 49, 32)	117,344
	Separable_Conv2D_6	(None, 64, 64, 256)	68,096		Point_Conv2D_8	(None, 16, 16, 256)	65,792
	Conv2D_7	(None, 64, 64, 256)	590,080		Point_Conv2D_9	(None, 8, 8, 256)	65,792
	Maxpooling2d_3	(None, 32, 32, 256)	0		Point_Conv2D_10	(None, 7, 7, 192)	196,800
B4	Separable_Conv2D_11	(None, 32, 32, 512)	133,888	A3	Multi-head attention_A_3	(None, 49, 32)	117,344
	Separable_Conv2D_12	(None, 32, 32, 512)	267,264		NormL_(123)	(None, 7, 7, 32)	64
	Conv2D_13	(None, 32, 32, 512)	2,359,808		Layers_Add	(None, 7, 7, 32)	0
	Maxpooling2d_4	(None, 16, 16, 512)	0		dense_13 (Dense)	(None, 4)	132
	Global_average_pooling2d_1	(None, 32)	0				

איור 32 - כמות הניורונים לפי בלוקים [12].

איור 33 מסכם את המבנה של כל אחד מהבלוקים המרכיבים את רשת הניורונים השלמה. ניתן לראות את גודל שכבות הקונבולוציה, כמות שכבות האיגום, ואת השימוש בשכבות הקשב. בסוף רשת הניורונים ניתן לראות כי קיימת שכבת FC האחראית על הסיווג הסופי של הקלטים מבין 4 סוגי המחלקות הקיימות במחקר.

ניתן לשים לב באיור כי ברשת הקונבולוציה קיימות שתי שכבות קונבולוציה שונות. סוג השכבה הראשונה היא שכבת D2CONV – שכבה זו מיישמת את פעולת הקונבולוציה הסטנדרטית. השכבה השנייה נקראת Depthwise separable convolution. או קונבולוציה ספרבילית עמוקה.

הרעיון המרכזי מאחורי קונבולוציה ספרבילית עמוקה היא פיצול פעולת הקונבולוציה לשני חלקים שונים – קונבולוציה הפועלת על מרחב התמונה וקונבולוציה הפועלת על עומק התמונה. קונבולוצית העומק פועלת על תא בודד במישור התמונה אך לעומק נרחב של מספר ערוצי הקלט (כלומר על פיקסל בודד לאורך כל ערוציו), וקונבולוצית המרחב היא קונבולוציה סטנדרטית הפועלת על מישור התמונה.

במחקר זה לשכבת הקשב יש תפקיד כפול. השכבה נמצאת בשימוש גם בכדי למקסם את היכולת להפיק מידע מהקלט וגם בכדי לנסות ולמזער את הכוח החישובי הדרוש לניתוח הקלט. השימוש בשכבות קשב ממקסם את היכולת לעבד את מירב המידע משכבות הקלט, לקראת המעבר לשכבות המבצעות סיווג ובכך לחסוך בכמות שכבות הקונבולוציה ושכבות האיגום. במקום בניית רשת קונבולוציה עמוקה, בעלת כמות גדולה של שכבות קונבולוציה – החוקרים ניסו לצמצם את כמות שכבות הקונבולוציה באמצעות שימוש במנגנון הקשב.

4.3.4 תוצאות המחקר ומסקנות

המודל שהוצג במחקר הגיע לרמת דיוק של 96.75 אחוז על קבוצת המבחן, רמת רגישות של 96.6%, וספציפיות של 98.06%.

Datasets	Methods	Train_loss	Train_acc (%)	Val_loss	Val_acc (%)	Size (M)	Specificity (%)	Sensitivity (%)	Test_acc (%)
Four classes (X-ray images)	LeNet-5	0.2418	94.72	0.3494	90.72	51.5	89.06	91.27	91.27
	AlexNet	0.2065	95.44	0.3059	90.08	444	92.13	90.13	90.15
	GoogLeNet	0.2832	92.40	0.2944	86.12	52	87.64	89.25	88.12
	ResNet	0.1931	97.15	0.2069	93.19	270	96.73	91.38	93.38
	VGGNet-16	0.1845	97.89	0.1891	92.40	114	94.52	92.05	92.45
	MGMADS-1	0.0602	98.63	0.0602	93.31	40.1	96.03	93.12	93.20
	MGMADS-2	0.0485	99.03	0.0485	94.87	41.8	96.12	95.12	95.12
	MGMADS-3	0.0139	99.62	0.0140	96.25	43.6	98.06	96.60	96.75

טבלה 6 - תוצאות המחקר [12].

בטבלה מעלה ניתן לראות את תוצאות המחקר על גבי תמונות הרנטגן של בית החזה בלבד (וללא שימוש בתמונות ה-CT שהופיעו במחקר כאמור). ממד השורות מציג רשתות שונות שנבחנו במחקר או במחקרי עבר כאשר חמשת השורות הראשונות הן רשתות קונבולוציה

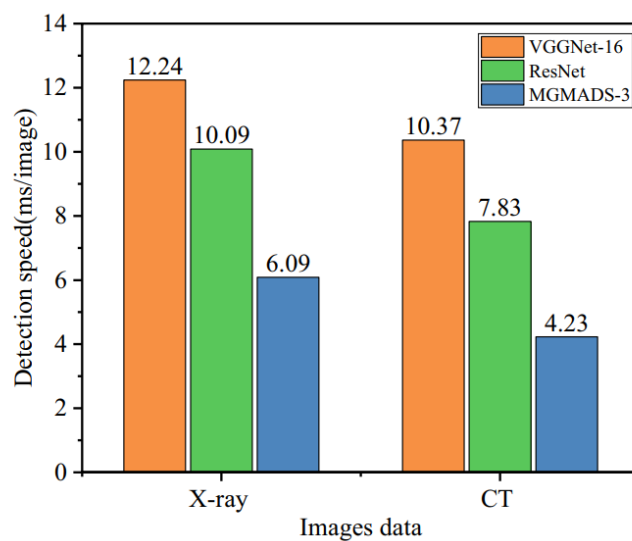
טהורות, ושלושת השורות האחרונות הן סוגים שונים של מודל הקונבולוציה משולב מנגנון קשב שפותח במחקר הנ"ל.

החוקרים מציגים את תוצאות פונקציית ההפסד ואת אחוז הדיוק על קבוצת האימון (עמודה 1-2), את ערכי פונקציית ההפסד ואחוז הדיוק על קבוצת הוולידציה (עמודות 3-4), ואת התוצאות הסופיות של המודל בהתאם למדד ה Accuracy על בסיס קבוצת המבחן בעמודות 6-8.

כאמור במחקר זה הושם דגש רב גם על ביצועים, קרי גודל המודל וזמן הריצה שלו. עמודה מספר 5 מציינת את גודל המודל במיליון בייטים.

ניתן לראות כי יחסית לרשתות קונבולוציה אחרות כגון Resnet (270 מיליון byte), או 16VGG (114 מיליון byte), גודל המודל שהוצע במחקר זה עומד על כ-40 מיליון byte בלבד.

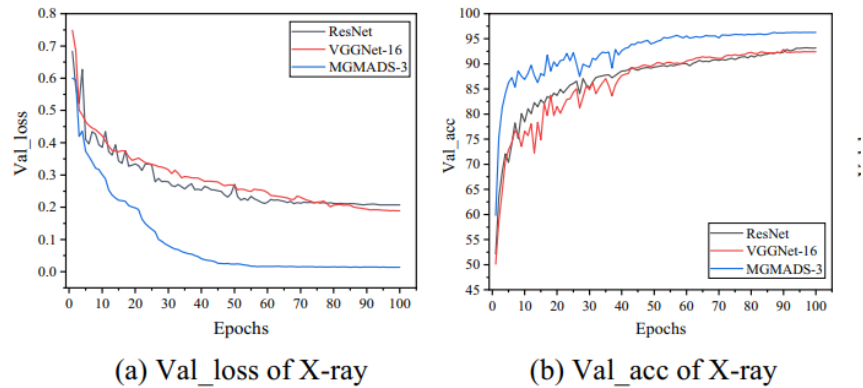
בהמשך לתהליך ניתוח הביצועים, מציגים החוקרים את ההבדל בזמן הריצה של המודלים השונים. לצורך כך ביצעו החוקרים השוואה בזמן הריצה בין המודל שפותח על ידם ובין רשתות נירונים אחרות ובפרט רשתות Resnet 16vvgi. ניתן להתרשם בהבדלים במהירות הריצה באיור מטה:



איור 33 - מהירות הריצה של מודלים שונים [12]

מהירות הריצה הינה ביחידות של מילי שניה לתמונה. ניתן לראות כי ישנו שוני מהותי בזמני הריצה בין המודלים השונים. בעוד שרשתות הקונבולוציה 16VGG ו-RESNET הציגו זמני ריצה ארוכים גם עבור מאגר המידע המכיל תמונות רנטגן וגם עבור מאגר המידע המכיל תמונות CT, ניתן לראות כי רשת הנירונים שפותחה במהלך המחקר מגיע לביצועים טובים יותר מבחינת זמני הריצה.

בנוסף מציגים החוקרים את התכנסות תוצאות פונקציית ההפסד ואת אחוז הדיוק של המודל שלהם במהלך תהליך הלימוד. הגרף מציג את תוצאות תהליך הלימוד של המודל (בקו הכחול) בהשוואה לשתי רשתות קונבולוציה מוכרות – Resnet (בירוק) ו16vgg (בכתום).



איור 34 - תיאור פונקציית ההפסד ורמת הדיוק בתהליך האימון. [12]

ניתן לראות הן בגרף השמאלי כי קצב ההתכנסות של רשת הניורונים המוצעת במחקר מהירה יותר מאשר קצב ההתכנסות של הרשתות 16VGG וRESNET. בנוסף, ערך פונקציית ההפסד עבור הרשת שנבנתה במחקר טובה יותר מאשר רשתות הקונבולוציה המתחרות. בגרף הימני ניתן לראות שגם עבור קבוצת הוולדציה, מקבלים תוצאות זהות, קרי קצב התכנסות מהיר יותר, ותוצאה סופית טובה יותר עבור רשת הניורונים שפותחה במהלך המחקר.

לסיכום – במחקר זה מוצגת רשת ניורונים מתקדמת המחברת שכבות קונבולוציה למנגנון הקשב. הרשת מביאה לביצועים מתקדמים לצד ניסיון לחסוך במשאבים (מקום וזמן ריצה).

4.4 סיווג מחלות ריאתיות באמצעות רשתות GAN-CNN

4.4.1 הקדמה

פרק זה יעסוק במקרה בוחן מעניין וייחודי המשלב בין רשת קונבולוציה עם רשת גנרטיבית. מטרת רשת הנורונים הנ"ל היא לשפר את תהליך הסיווג של תמונות הרנטגן למחלות הריאה השונות באמצעות התמקדות באזורי עניין משמעותיים בריאות במטרה לשפר את דיוק רשת הנורונים

[1] מעידים במחקרם כי במשימת סיווג תמונות הרנטגן באמצעות כלי למידה עמוקה כמעט ולא שמים דגש על עיבוד מקדים של תמונות הקלט במטרה למקסם ולתעל אזורי עניין משמעותיים בתמונת הקלט ומתמקדים בעיקר בנרמול הקלט בצורות שונות. החוקרים טוענים כי מיקוד בנקודות עניין (ROI) בבית הריאה טרום שלב הזנת התמונות לרשת הנורונים עצמה הוא משמעותי עבור הליך הסיווג.

החוקרים השתמשו ברשתות נורונים גנרטיביות (GAN) בכדי למקד את תמונות הקלט בריאות עצמן ולא בבית החזה כולו. בנוסף מנסים החוקרים להשתמש בכלי ראייה ממוחשבת קלאסיים כגון SIFT או BRISK כדי לחלץ מהריאות נקודות עניין משמעותיות במקביל להזנת תמונות הקלט לרשת נורונים קונבולוציונית.

SIFT ו-BRISK הינם אלגוריתמים מתחום הראייה הממוחשבת אשר מטרתם היא לזהות נקודות ייחודיות בתמונות. אלגוריתמים אלו מהווים את הבסיס לטכניקות שונות לביצוע רגיסטרציה בין תמונות.

4.4.2 מטרות המחקר

מטרת המחקר היא לנסות ולטייב את תמונות הקלט ע"י שימוש ברשתות נורונים גנרטיביות וכלי ראייה ממוחשבת קלאסיים. את התמונות שעברו עיבוד מקדים בצורה זו מזרימים לאחר מכן לרשת קונבולוציה שאומנה באמצעות למידת העברה. בפרט משתמשים ברשתות נורונים מסוג VGG16, DENSNET 201 וכו'.

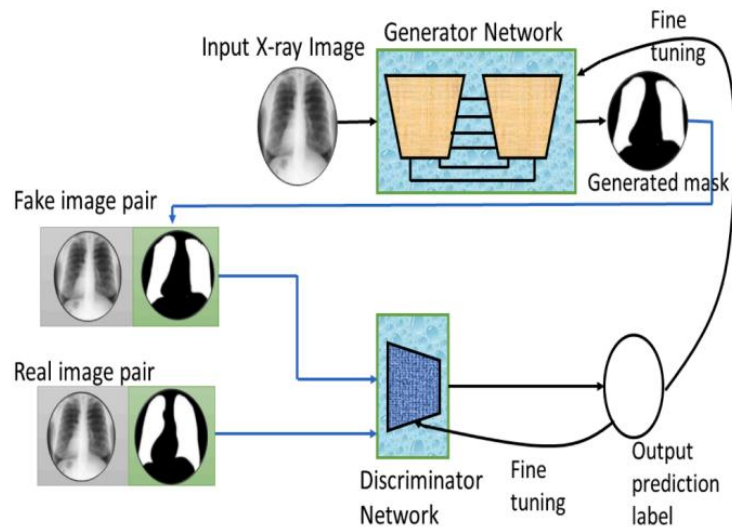
המחקר מתמקד בניחוח תמונות משלוש מחלקות – תמונות של אנשים בריאים, תמונות של חולי קורונה ותמונות של חולים בדלקת ריאות.

4.4.3 מאגר הנתונים

הנתונים במחקר זה מחולקים לשני מאגרי מידע – מאגר המידע שבו נעשה שימוש לטובת אימון הרשת הגנרטיבית ומאגר הנתונים שבו נעשה שימוש ללימוד רשת הנורונים המסווגת. עבור אימון הרשת הגנרטיבית נעשה שימוש ב-247 תמונות בגודל 224/224/3 בגווי RGB. סט נתונים זה זה חולק לסט אימון, אימות ומבחן אך בשל גודלו סט המבחן כלל רק כ-5% מסך התמונות, וסט האימות כלל כ-10% מסך התמונות. בשל הגודל הקטן של סט הנתונים בוצעו אוגמנטציות לקלטי המקור. האוגמנטציות התבססו על התמרות אפניות בלבד. עבור רשת הסיווג בוצע שימוש בסט נתונים שונה. סט הנתונים הורכב מ-930 תמונות של אנשים החולים בקורונה, 1583 תמונות של אנשים בריאים, ו-4273 תמונות של אנשים החולים בדלקת ריאות. בסט הנתונים הזה בוצע תהליך של אוגמנטציה ע"י שימוש בהתמרות אפניות.

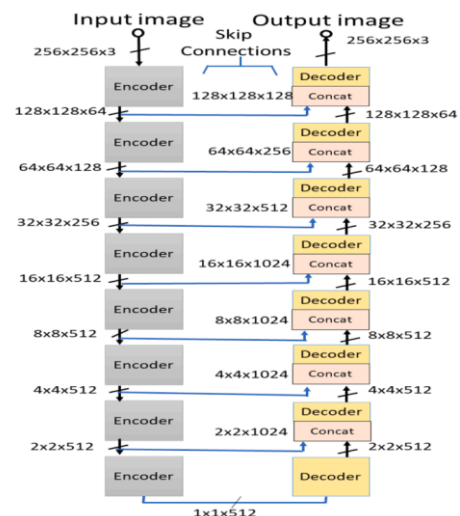
4.4.4 מהלך המחקר

אימון הרשת הגנרטיבית – מטרת הרשת הגנרטיבית היא כאמור לאפשר לרשת הסיווג להתמקד בתמונות קלט הכוללות את הריאות עצמן בלבד. לשם כך אומנה רשת גנרטיבית. המחולל (גנרטור) קיבל כקלט צילומי רנטגן של בית הריאות וכפלט נדרש להוציא תמונת "מסיכה" של הריאות עצמן. כלומר תמונה המסמנת את מיקום הריאות בלבד בתוך תמונת הקלט. המפלה (דיסקרימינטור) קיבל כקלט זוג תמונות – תמונת רנטגן ומסיכה תואמת ונדרש לקבוע האם התמונות יוצרו בצורה מלאכותית על ידי המחולל או שהן תמונות אמיתיות. תהליך האימון האיטירטיבי מביא לכך שתמונות המסכה של המחולל מגיעות לאיכות גבוהה כך שניתן להוציא מתמונות המקור את החיתוך המדויק של הריאות מתוך תמונת בית החזה.



איור 35 - תהליך אימון הרשת הגנרטיבית [1]

גם תת הרשת המרכיבה את המפלה וגם תת הרשת המרכיבה את המחולל התבססו על למידת העברה המבוססת על מודל בשם CGAN PIX2PIX.

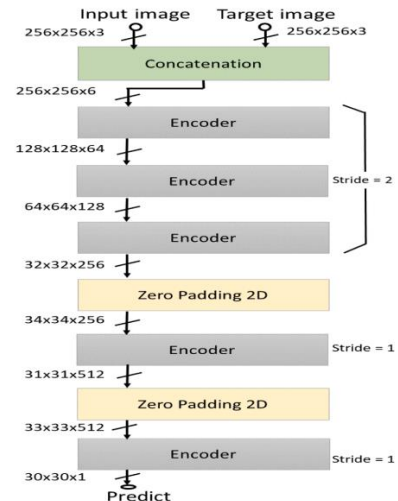


איור 36 - ארכיטקטורת המחולל [1]

המחולל נבנה על בסיס ארכיטקטורה של מקודד-מפרש, המבוססת על מבנה U, הוא מורכב משכבות של תאי קידוד המוערמים זה על גבי זה במטרה לדחוס את תמונת הקלט ולחלץ את המאפיינים בתוכה. תאי הקידוד הם למעשה תאים המורכבים משכבות קונבולוציה.

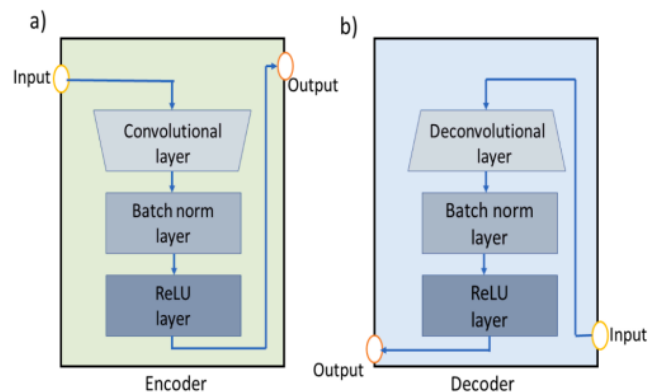
לאחר תהליך הדחיסה מועבר ווקטור נתונים לתוך רכיב המפרש, אשר מבצע פירוש בחזרה בין תאי פירוש שונים. תאי הקידוד והפירוש מחוברים יחדיו כפי שמודגם בשרטוט מעלה, ומידע על תמונת הקלט עובר ישירות בין הרמות השונות של תאי הקידוד והפירוש. במילים אחרות כל תא של המפרש מקבל כקלט את הפלט של תא המפרש הקודם לו מחד וגם את

הפלט של תא הקידוד המקביל אליו. תאי המפרש הם תאים המורכבים משכבות קונבולוציה הפוכות, אשר מגדילות את ממדי הקלט באמצעות גרעין הקונבולוציה במקום להקטין.



איור 37 - ארכיטקטורת המפלה [1]

המפלה נבנה על בסיס ארכיטקטורה של מפרש מסוג GAN-PATH, ובנוי משכבות של תאים המבצעים קידוד. קלט המפלה הוא 2 תמונות קלט – מסיכה ותמונת רנטגן של בית הריאות משורשרות יחדיו. הפלט של המפלה כאמור הוא תוצאת סיווג – האם מקור התמונות הוא מלאכותי ונוצר על ידי המחולל או שמקור התמונה אמיתי.



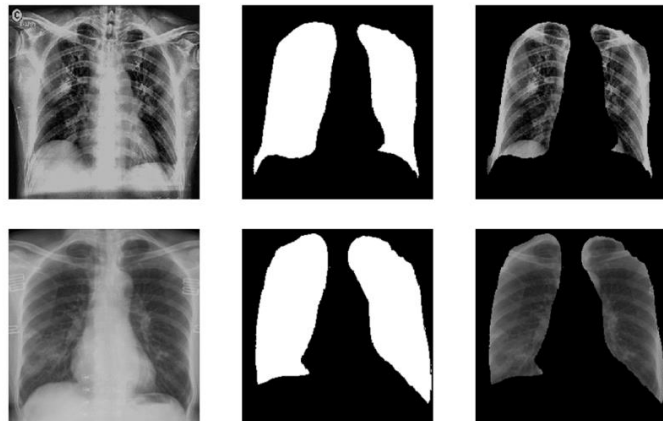
איור 38 - תאי קידוד ופירוש [1]

נדון כעת במבנה הפנימי של תאי הקידוד והפירוש שנמצאים בתוך ארכיטקטורת המחולל והמפלה.

תאי הקידוד הם למעשה תאים המורכבים משכבות של רשתות נוירונים. תא המקודד מצד שמאל באיור מעלה מורכב משלוש שכבות נוירונים – שכבת קונבולוציה, שכבת נרמול

ושכבת FC. תא הפירוש מצד ימין באיור מעלה מורכב מאותן שכבות בסדר הפוך כאשר במקום פעולת הקונבולוציה מבוצעת פעולת דיקונבולוציה.

פלט הרשת הגנרטיבית הוא "מסיכה" המציינת את מיקום הריאות בתמונת הקלט. באמצעות מסיכה זו ניתן לבדוד את הריאות עצמן מתמונות המקור בכדי לבצע עליהן תהליכים נוספים שסופם סיווג תמונת הקלט בין מחלקות הסיווג השונות.



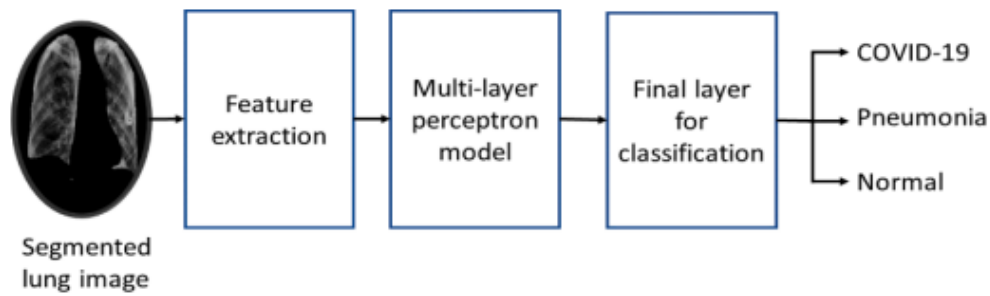
איור 39 - פלט רשת GAN [1]

באיור מעלה ניתן לראות מצד שמאל תמונת קלט של בית החזה. במרכז האיור ניתן לראות את המסכה המתאימה שיוצרה על ידי הרשת הגנרטיבית ומצד ימין את חיתוך הריאות באמצעות המסכה מתמונת הקלט. תמונת הריאות בלבד היא הקלט של רשת הקונבולוציה.

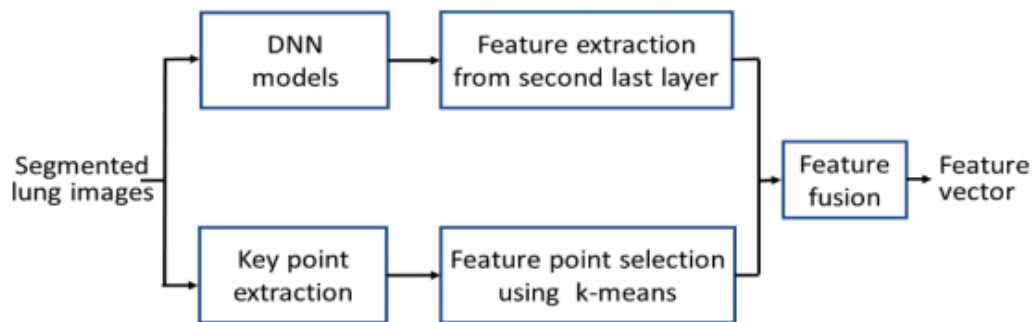
רשת הסיווג ותהליך הוצאת פיצ'רים – הוצאת פיצ'רים נחשב לאחד מהשלבים המרכזיים בתהליך סיווג תמונות באמצעות כלי למידה עמוקה. בעוד שבעבר בוצע שימוש נרחב בכלי ראייה ממוחשבת ועיבוד תמונה בכדי לבצע את התהליך בשנים האחרונות כלי למידה עמוקה נמצאו בשימוש נרחב גם בתהליך הוצאת הפיצ'רים.

מחקר זה מנסה להשתמש בשילוב של כלי ראייה ממוחשבת קלאסיים לצד שימוש ברשתות נוירונים בכדי לשפר ולטייב את תהליך הוצאת הפיצ'רים מתמונות הקלט.

לאחר הכנת התמונות ע"י הרשת הגנרטיבית, התמונות מוזרמות לתוך שכבה שמטרתה הוצאת פיצ'רים. שלב הוצאת הפיצ'רים כולל שימוש במקביל ברשת קונבולוציה לצד כלי ראייה ממוחשבת קלאסיים כגון SIFT. לאחר שהתמונות עוברות דרך כלי הוצאת הפיצ'רים השונים הן מוזרמות בהמשך לראש סיווג. כך למעשה התמונות שעברו סגמנטציה מועברות בכלי עיבוד שונים במקביל לטובת טיוב תהליך הוצאת הפיצ'רים.



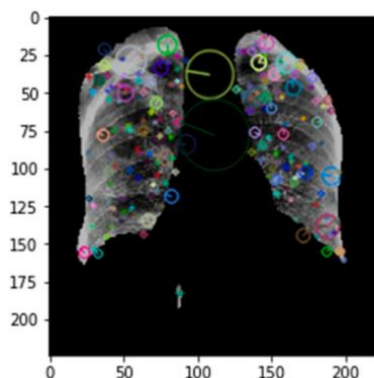
איור 40 - מבנה המערכת בשלב הוצאת הפיצ'רים [1]



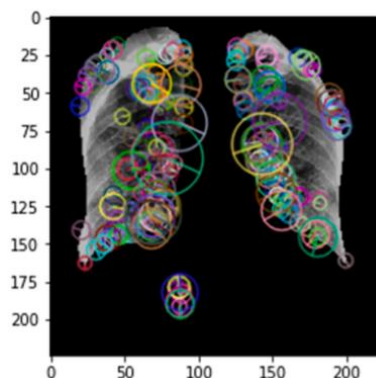
איור 41 - תהליך הוצאת הפיצ'רים ברשת [1]

במחקר בוצע שימוש בלמידת העברה לטובת השימוש ברשתות קונבולוציה. המחקר ביצע השוואה בין רשתות קונבולוציה שונות כגון VGG16/19, ו DENSNET169/201. בנוסף ניסו החוקרים לפתח רשת קונבולוציה עצמאית. אלגוריתמי הראייה הממוחשבת הנוספים בהם נעשה שימוש הם SIFT ו BRISK.

a) SIFT



b) BRISK



איור 42 - וויזואליזציה של תוצרי אלגוריתם SIFT ו BRISK [1]

בסוף תהליך הוצאת הפיצ'רים, מוזרמות התמונות לראש סיווג המורכב משלוש שכבות FC בעלות גדלים משתנים – שכבת הסיווג הראשונה בנויה מ-4096 נירונים, השכבה השנייה מ-1024 נירונים והשלישית מ-256 נירונים. בסוף ראש הסיווג נמצאת שכבת סיווג אחרונה בעלת 3 נירונים, עם פונקציית אקטיבציה שונות כגון SVM, RF, SOFTMAX וראשי סיווג נוספים.

4.4.5 תוצאות המחקר ומסקנות

החוקרים מדגישים מספר כלי הערכה לבחינת תוצאות המחקר. ראשית בוחנים החוקרים את ביצועי המודל מבחינת זמני הריצה, בדגש על ביצועי רשתות הקונבולוציה. רשתות הקונבולוציה השונות הן בעלות ביצועי זמן עיבוד שונים. התוצאות מוצגות יחסית לזמן אימון הרשתות, ויחסית לזמן האימון של אפוק בודד.

	Training time	Average time per step
VGG-16	1046s	174.3 ms
VGG-19	1289s	214.8 ms
DenseNet-169	811s	135.1 ms
DenseNet-201	2298s	383.0 ms
sCNN	200s	33.3 ms

טבלה 7 - זמן האימון של רשתות הקונבולוציה [1]

זמן האימון המהיר ביותר הוא של הרשת העצמאית שפיתחו החוקרים והוא קטן משמעותית יחסית לזמני הריצה של רשתות הקונבולוציה האחרות.

בנוסף לכך מבצעים החוקרים השוואה בין כלל השילובים האפשריים בתהליך הוצאת הפיצ'רים, קרי שילוב של SIFT ו BRISK עם רשתות הקונבולוציה השונות, לצד שימוש ברשתות קונבולוציה בלבד. ההשוואה בביצועים מתבססת על מדד ההערכה של ACCUACY לצד מדדי הערכה נוספים.

Model & Feature extraction	Final layer	Avg Acc (%)	Avg Sens (%)	Avg Spec (%)	Avg Prec (%)	Avg F1-score (%)
BRISK VGG-16	Softmax	95.9	93.9	96.9	93.9	93.9
	SVM	95.9	93.9	96.9	93.9	93.9
	RF	95.9	93.8	96.9	93.9	93.9
	XG-Boost	95.6	93.6	96.8	93.7	93.5
BRISK VGG-19	Softmax	96.5	94.6	97.2	94.6	94.6
	SVM	96.4	94.6	97.2	94.6	94.6
	RF	96.6	95.0	97.4	95.0	95.0
	XG-Boost	96.1	94.4	97.1	94.2	94.3
BRISK sCNN	Softmax	94.7	92.3	96.0	92.1	92.1
	SVM	95.2	93.0	96.4	92.8	92.8
	RF	95.4	93.3	96.5	93.1	93.2
	XG-Boost	95.4	93.3	96.5	93.1	93.2
BRISK DenseNet 169	Softmax	95.2	92.9	96.5	93.0	92.7
	SVM	94.9	92.5	96.2	92.5	92.5
	RF	95.2	92.8	96.4	93.0	92.8
	XG-Boost	95.2	92.9	96.5	93.0	92.7
BRISK DenseNet 201	Softmax	94.9	92.4	96.2	92.5	92.4
	SVM	94.7	92.2	96.0	92.1	92.1
	RF	94.5	91.8	95.8	91.8	91.8
	XG-Boost	94.5	91.8	95.8	91.8	91.8

טבלה 8 - תוצאות המחקר [1]

בטבלה מעלה ניתן לראות את פירוט התוצאות הטובות ביותר אליהם הגיעו החוקרים. הטבלה מחולקת ל-5 מודלים ראשיים שהביאו את התוצאות הטובות ביותר – מודל BRISK המשולב עם רשתות קונבולוציה שונות (עמודה 1). עבור כל אחת מהארכיטקטורות הללו, מוצגות התוצאות על פי חלוקה לראשי הסיווג השונים שנבחנו (עמודה 2).

עמודה 3 מציגה את ציון Accuracy, שמהווה את מדד ההערכה המרכזי במחקר, ועמודות 4-7 מציגות מדדי הערכה נוספים. ניתן לראות כי אחוז הדיוק הטוב ביותר הגיע משימוש ברשת 19VGG בשילוב עם ראש סיווג מסוג RF ועומד על 96.6%.

לסיכום, מחקר זה מציג גישה ייחודית של שילוב רשת גנרטיבית לצד רשת סיווג קונבולוציונית. רשת הקונבולוציה ממומשת במחקר זה על ידי למידת העברה. בנוסף, במהלך המחקר בוצע מחקר והשוואה בין ראשי סיווג שונים.

5. מדדי הערכה לניתוח השוואתי

ישנם מספר מדדי הערכה נפוצים אשר נמצאים בשימוש כאשר אנו מנסים למדוד ולהעריך יעילות ודיוק של רשתות נוירונים בביצוע סיווג של מידע. בפרק זה אסקור מספר מדדי הערכה שונים הנמצאים בשימוש נפוץ במחקר ואעמוד על המשמעות והיישום שלהם.

מדדי הערכה ממלאים תפקיד מרכזי במדידת האפקטיביות והיעילות של מודלי למידה עמוקה. המדדים מספקים מדדים כמותיים עבור ביצועי המודל ומאפשרים לחוקרים להשוות מודלים שונים, לכוון פרמטרים ולקבל החלטות מושכלות לגבי בחירת הארכיטקטורה המיטבית לסוגיה הנבדקת. בעולם הלמידה עמוקה, שבה רשתות עצביות מורכבות מאומנות על מערכי נתונים גדולים מאוד יש למדדי הערכה חשיבות מכרעת בניתוח רשתות הנוירונים וביצוען [8].

ראשית אבחין בין מספר מושגים בסיסיים הנוגעים להערכת מודלים. כאשר אנו חוקרים את ביצועיו של מודל אנו מעוניינים לעמוד על יכולתו של המודל לסווג נכון בין תצפיות הלקוחות מתוך מאגר מידע. בפרט אנו מעוניינים לדעת אם המודל מצליח לסווג תצפית "חיובית" בתור תצפית חיובית, ותצפית "שלילית" בתור תצפית שלילית. לדוגמה, במידה ורשת הנוירונים לומדת מאגר תמונות של כלבים וחתולים ומטרת הרשת היא לסווג נכון תמונות אלו מטרתנו היא לוודא שמחד כל תמונה של חתול תתויג על ידי המודל כחתול, אך גם שכל תמונה של כלב תתויג שכזו. אין אנו מסתפקים בכך שרק סוג מחלקה אחד יתויג בצורה נכונה.

כדי לדון במקרה הרחב של תהליך הסיווג נניח כי מאגר המידע שברשותנו מכיל דוגמאות חיוביות ושליליות בסיווג. מטרתנו היא לבנות מודל המצליח לסווג כל תמונת קלט בין שתי המחלקות. בכדי לעמוד על שוני זה אנו מגדירים מספר מצבים בהם מתקיימת או לא מתקיימת קורלציה בין הערך האמיתי של התצפית לבין החיזוי של המודל [27]:

TP – מצב בו התצפית היא חיובית והמודל סיווג את התצפית כחיובית.

FN – מצב בו התצפית היא חיובית, אך המודל סיווג אותה כשלילית.

FP – מצב בו התצפית היא שלילית, אך המודל סיווג אותה כחיובית.

TN – מצב בו התצפית היא שלילית והמודל סיווג אותה כשלילית.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

טבלה 9 - טבלת סיכום חיזוי מול מצב אמיתי של תצפית.

נשים לב כי המודל טועה לגבי הסיווג בקבוצת FP ובקבוצת FN בלבד. שהרי בקבוצת TP ובקבוצת TN תוצאות סיווג המודל שווה לתוצאות הסיווג האמיתיות של הדוגמה.

קבוצות אלו והיחס ביניהן מגדירות מספר מדדי הערכה מרכזיים, אשר עוזרים להעריך את נכונות ויעילות ביצועי המודל. מדדים אלו נפוצים בשימוש ומיועדים להשוואת רשתות נוירונים. כל מדד נותן מענה לסוגייה שונה במעט ומדגיש צד אחר בהערכת הנתונים. בכדי לקבל תמונה מלאה על דיוק ואיכות המודל הנבחן ישנו צורך להשתמש ביותר ממד הערכה אחד באותו זמן.

5.1 מדד Accuracy

מדד ההערכה הנפוץ בתחום הערכת רשתות נוירונים הוא מדד הדיוק (Accuracy). המדד בודק את היחס שבין כמות החיזויים הנכונה של המודל (בין אם המודל חזה סיווג חיובי עבור דוגמה חיובית ובין אם המודל חזה סיווג שלילי עבור דוגמה שלילית) יחסית לכלל החיזויים שביצע המודל. מדד זה נפוץ כיוון שהוא אינטואיטיבי להבנה ומציג בסופו של דבר את "תוצאה הסופית" – כמה חיזויים ביצע המודל נכון יחסית לכמות החיזויים שבוצעו.

את מדד הדיוק ניתן לחשב בצורה הבאה:

$$Accuracy = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

משוואה 13 - [27].

בעוד שהדיוק אינטואיטיבי וקל להבנה, יש לו גם חסרונות. ערך הדיוק לא מתאים למערכי נתונים לא מאוזנים, שבהם מחלקה אחת נפוצה הרבה יותר מאשר אחרות. במקרים כאלה, ערך דיוק גבוה עשוי להסוות את הביצועים הגרועים של המודל עבור דוגמאות בעלות ערך תיוג שאינן נפוצות ובכך ליצור תמונת שווא של דיוק גבוה למודל כולו בזמן שרק הקבוצה המרכזית מתויגת בצורה נכונה [27].

לדוגמה, אם מאגר המידע שלנו מכיל תמונות רבות חיוביות ומספר מועט של תמונות שליליות, ונניח כי המודל מסווג את כל התמונות במאגר המידע כתמונות חיוביות. אחוז הדיוק יכול להיות עדיין גבוה וזאת על אף שהמודל "מסתיר" בתוכו טעות יסודית בנוגע למחלקת התמונות השליליות. צורת התנהגות זו של המודל איננה משקפת אחוז דיוק גבוה אמיתי וניתן להסיק מכך שהמודל מוטה לטובת סיווג כלל התמונות כתמונות חיוביות.

5.2 מדד Precision

מדד הערכה נוסף הוא מדד Precision. מדד זה מתמקד בנכונות התחזיות "החיוביות". המדד מוגדר כיחס בין התחזיות החיוביות שהן באמת חיוביות יחסית לכמות התחזיות שתיוגן כחיוביות, גם אם הן בפועל הן לא חיוביות. המדד מוגדר בצורה הבאה:

$$precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

משוואה 14 - [27].

המדד מנסה לכמת את הדיוק של התחזיות החיוביות שנעשו על ידי המודל. השימוש בו עונה על השאלה: "מכל המקרים שנחזו כחיוביים, כמה היו באמת חיוביים?" ערך Precision גבוה מצביע על כך שהמודל מבצע תחזיות חיוביות ברמת דיוק גבוהה. השימוש במדד זה הוא טוב מאוד כאשר המחיר של תיוג FP הוא קריטי לדוגמה כאשר עוסקים בדאטה רפואי. תיוג חולה כ-FP עלול לגרום למתן טיפול שאיננו נחוץ ואף לגרום לפגיעה בחולה. המיקוד במדד תורם לאמינות החיזויים החיוביים ומקטין את הסיכוי לתיוגים מסוג FP.

המדד בא לידי ביטוי חיובי בהתמודדות עם מערכי נתונים לא מאוזנים. במקרים כאלה Precision מספק תובנות לגבי ביצועי המודל עבור מחלקות שנמצאות בשכיחות נמוכה במרחב המדגם [27].

5.3 מדד Sensitivity

מדד הרגישות, המוכר גם בשם recall הוא מדד הערכה נוסף. רגישות מודדת את היכולת של מודל לזהות נכון מקרים חיוביים מתוך כל המקרים החיוביים בפועל. הוא מוגדר כיחס של תחזיות חיוביות אמיתיות למספר הכולל של מקרים חיוביים בפועל. ניתן לחשב את המדד בצורה הבאה:

$$Sensitivity = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

משוואה 15 - [27].

מדד הרגישות חשוב במיוחד כאשר מחיר הטעות של אי סיווג דוגמה כחיובית הוא גדול. למשל בהליך רפואי – FN מתורגם לחוסר זיהוי של תופעה או מחלה אצל החולה על אף שהמחלה קיימת בגופו.

במילים אחרות, כאשר זיהוי חיובי הוא משמעותי מאוד ויש צורך להימנע מזיהוי לא נכון (גם במחיר של זיהוי סרק) יש צורך לשמור על רמת רגישות גבוהה [27].

5.4 מדד Specificity

מדד ה Specificity מודד את היכולת של המודל לסווג דוגמאות שליליות. המדד מחושב בצורה הבאה:

$$Specificity = \frac{True\ Negatives\ (TN)}{True\ Negatives\ (TN) + False\ Positives\ (FP)}$$

משוואה 16 - [27].

כאשר מדד זה מקבל ערך גבוה הדבר מעיד על כך שרשת הנוירונים מצליחה לסווג בצורה טובה דוגמאות שליליות ונמנעת מסיווג FP, ערך נמוך של המדד מעיד על כך שהמודל לא מצליח לסווג נכון דוגמאות שליליות, ויכול לגרום להתראות שווא רבות.

ניתן לומר ש Specificity היא מדד משלים למדד הרגישות (Sensitivity), שני המדדים יחדיו מספקים מידע לגבי יכולתו לסווג דוגמאות חיוביות ושליליות בצורה נכונה. בהקשר למימד הרפואי מדד זה משמעותי כאשר אנו רוצים לוודא שאנשים בריאים אינם מסווגים בטעות כחולים ובכך לצמצם התראות שווא בדבר מחלות.

בנוסף בין שני המדדים לרוב יש קשר של יחסי גומלין ולעיתים שיפור בממד אחד יכול לבוא על חשבון הממד האחר.

5.5 עקומת ROC

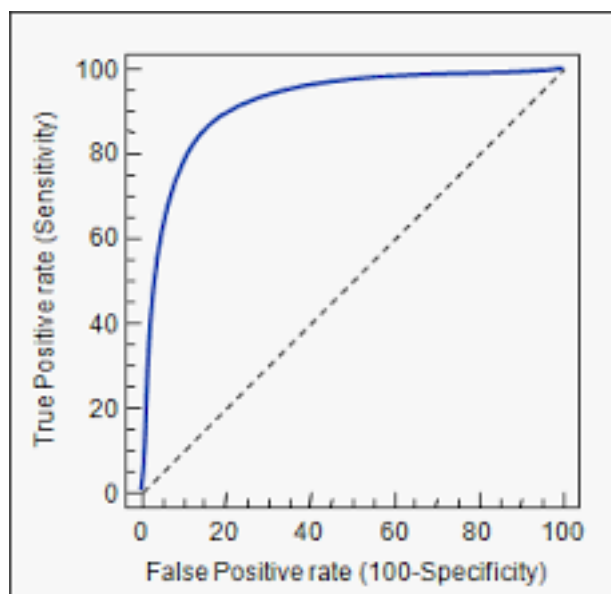
עקומת ROC - Receiver Operating Characteristic הינה עקומה גרפית למודל סיווג בינארי המציגה את הקורלציה בין אחוזי TP וה-FP.

השטח מתחת לעקומת ROC הוא ערך סקלרי המכמת את ביצועי המודל. ככול שערך זה גבוה יותר ל-1 (100%) כך המודל מבצע פרידיקציה טובה יותר (עד כדי מושלמת) על קבוצת המדגם. למודל בעל יכולת סיווג טובה יותר תהיה עקומת ROC העולה בחדות מהקצה השמאלי תחתון, ותמשך גבוה עד הפינה הימנית של העקומה – דבר שיעיד על אחוזי סיווג גבוהים.

קו האלכסון בעקומה מייצג מודל המבצע ניחוש אקראי, ולכן עקומות שעוברות מתחת לקו זה נחשבות לגרועות מאוד כיוון שהן מייצגות מודל הגרוע יותר מביצוע סיווג אקראי. עקומות ROC הינן כלי וויזואלי טוב להערכת והשוואת מודלים שונים.

שרטוט העקומה ע"י שרטוט אחוז הדוגמאות החיוביות שסווגו כחיוביות – True positive rate (TPR) יחסית לכמות הדוגמאות השליליות שסווגו כחיוביות – False positive rate (FPR).

כל נקודה על העקומה מייצגת ערך סף שנקבע על ידי המודל בכדי לבצע סיווג. העקומה מייצגת את יחסי הגומלין בין ערך ה-TPR לערך ה-FPR.



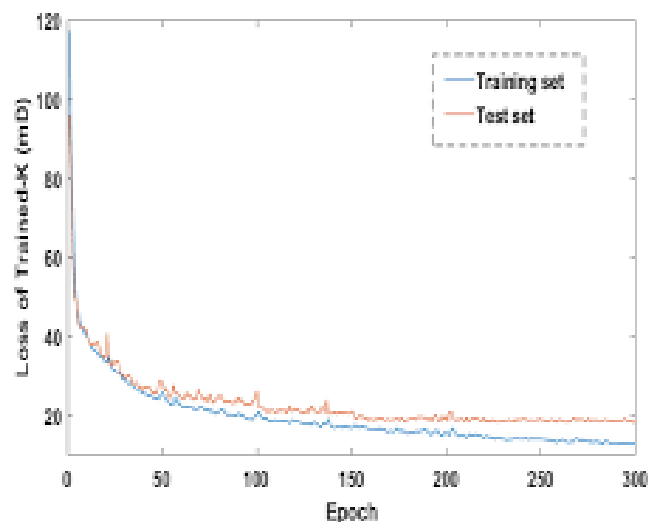
איור 43 - אילוסטרציה של עקומת ROC

5.6 פונקציית ההפסד LOSS

נשים לב כי כלל מדדי ההערכה שהוצגו עד כה מתמקדים בערך הסיווג הסופי של המודל כלומר כיצד ביצע המודל את הסיווג הסופי עבור דוגמה מסוימת. ערכי פונקציית ההפסד מייצגים פרטי מידע חשובים נוספים היכולים לשמש אותנו להערכת ביצועי מודל.

פונקציית ההפסד היא מרכיב קריטי באימון מודלים של למידה עמוקה. המטרה העיקרית של פונקציית הפסד היא לכמת את ההבדל בין הטעות של סיווג המודל לבין ערכי הסיווג האמיתיים. המטרה במהלך האימון של המודל היא למזער את ערכי פונקציית ההפסד.

הפונקציה עוזרת לכמת את איכות המודל, ולבצע אופטימיזציה במהלך האימון. על אף שפונקציית ההפסד לכשעצמה היא איננה מדד הערכה מרכזי בהערכת תוצאות מודלים, ולרוב משתמשים בה כדי לבחון את תהליך האימון – מחקר הפונקציה יכול להעיד על ערך התוצאות שהתקבלו ולספק כלי נוסף לבחינת טיבו של מודל.



איור 44 – אילוסטרציה של פונקציית הפסד

כאשר ערך פונקציית ההפסד גבוה הדבר אומר כי למודל יש שיעור חוסר ביטחון גבוה בנוגע לאיכות הסיווג שביצע. כאשר ערכי פונקציית ההפסד נמוכים הדבר מעיד כי רמת הביטחון של המודל בסיווג אותו ביצע היא גבוהה. נתון זה עוזר לנו ללמוד על רמת הביטחון של המודל בסיווגיו – ניקח לדוגמה תוצאות של שני מודלים בעלי ערכי דיוק (Accuracy) זהים. ערכים זהים מעידים על כך שהמודל סיווג נכון את אותה הכמות של דוגמאות. נניח כי למודל הראשון יש ערכים גבוהים של פונקציית ההפסד ולמודל השני ערכים נמוכים של הפונקציה. עובדה זו תעיד על כך שאומנם תוצאות הדיוק הן זהות בין המודלים, אך רמת הביטחון של המודל השני בחיזוייו טובות יותר. ולכן נשאף לעבוד עם מודל זה.

ערכי פונקציית ההפסד נדגמים לרוב גם לאורך תהליך האימון לאורך. אנו מצפים לזהות ירידה בערכי פונקציית ההפסד לאורך התקדמות האיטרציות של תהליך האימון. בנוסף ניתן להשוות את ערך פונקציית ההפסד על קבוצת המבחן ביחס לערך פונקציית ההפסד על קבוצת הוולידציה במהלך האימון. בחינה זו מאפשרת לנתח את קצב התקדמות לימוד המודל.

6. ניתוח השוואתי של השיטות המוצגות

קיים קושי להשתמש בכלל מדדי ההערכה שנסקרו בפרק 5 וזאת בשל העובדה כי לא כל המחקרים הציגו את כללי מדדי ההערכה הנדונים. חלק מהמחקרים העדיפו התמקדו במדד אחד בלבד ולא חשפו את תוצאות שאר מדדי הערכה.

נתון נוסף המקשה על ההשוואה בין המחקרים השונים נובע ממרחב הדגימה בהם השתמשו ומכמות המחלקות שסיווגו. מרחבי דגימה בגדלים שונים מאפשרים מצד אחד לזהות דפוסים רבים יותר בתהליך הלימוד של רשת הנורונים מאידך ניתן להניח שככול שמרחב המדגם גדול יותר כך יהיה קשה יותר לרשת הנורונים לזהות דפוסים שונים וייחודים במאגר הנתונים. כך שגודל המדגם שעליו אומן ובאמצעותו נבחנה רשת הנורונים הוא בעל השפעות אמביוולנטיות על ביצועי המחקר. ניתן להניח בזהירות שעל אף האמור לעיל, ככול שגודל מרחב הנתונים גדול יותר, כך משימת הסיווג מורכבת יותר.

בנוסף כמות המחלקות הנדרשות לסיווג משתנה בין המחקרים השונים. בעוד שכלל המחקרים ביצעו סיווג בין אנשים בריאים ובין אנשים החולים בקורונה, ישנם מחקרים שביצעו סיווג גם עבור מחלות נוספות וביניהן עכירות ריאה ודלקת ריאות. ככול שכמות מחלקות הסיווג גדולה יותר כך הדבר מקשה על ביצועי הרשת.

מאגרי המידע שעליהם נבחנו הרשתות המוצגות במחקר שונים זה מזה גם במקורות המידע. המחקרים ניתחו מקורות מידע שהגיעו ממקומות שונים ולכן גם נתון זה מקשה על יכולת ההשוואה שכן הקלטים של הרשתות השונות אינם זהים.

על אף האמור לעיל, נבצע השוואה של השיטות שנסקרו בעבודה באמצעות מדדי הערכה שהוצגו בפרק חמש. בנוסף ניתן את דעתנו לגבי השוני בכמות המחלקות שסווגו בכל מחקר וגודל קבוצת הנתונים עליהם אומן ונבחן המודל.

להלן סיכום ממצאי המדדים של ארבעת השיטות שנסקרו:

שם המחקר	סוג רשת	גודל המדגם	כמות מחלקות לסיווג	Accuracy	Specificity	sensitivity	F1-score
[5]	קונבולוציה טהורה	380	2	94.7	ללא	ללא	ללא
[31]	קונבולוציה משולבת LSTM	4575	4	99.2	99.8	98	98.8
[12]	קונבולוציה משולבת מנגנון קשב	17439	4	96.75	98.06	96.6	ללא
[1]	קונבולוציה ורשת גנרטיבית	6786 + 240	3	96.5	97.2	94.6	94.6

טבלה 10 - השוואת שיטות הסיווג.

מחקרם של [5] שבצע סיווג עם רשתות קונבולוציה בלבד הגיעה לתוצאות הדיוק (Accuracy) הנחותות ביותר מבין שאר שיטות הסיווג. מחקר זה לא הציג נתונים לגבי שאר מדדי הערכה. בנוסף שיטת מחקר זו התבססה על גודל מדגם קטן יחסית לשאר השיטות שנסקרו וכך גם כמות המחלקות שסווגה על ידי מחקר זה.

תוצאות רשת הקונבולוציה המשולבת עם מנגנון LSTM של [31] מגיע לתוצאות הדיוק (Accuracy) הטובות ביותר יחסית לשאר השיטות שהוצגו בעבודה. איכות התוצאות באה לידי ביטוי גם בשאר מדדי ההערכה שנבחנו. המחקר בוחן ארבע מחלקות סיווג שונות, וגודל המדגם במחקר הוא בסדר גודל בינוני.

רשת הקונבולוציה המשולבת עם רשת גנרטיבית של [1] הגיעה לתוצאות דיוק של 96.5% (Accuracy) על קבוצת המדגם. המחקר בחן קבוצת מדגם גדולה יחסית לשאר השיטות שנבחנו, אך כמות המחלקות שסווגו במחקר עומד על שלוש בלבד.

רשת הקונבולוציה המשולבת עם מנגנון הקשב של [12] היא בעלת קבוצת המדגם הגדולה ביותר מבין שאר שיטות הסיווג בסדרי גודל. בנוסף כמות המחלקות שסווגו במחקר עומד על ארבע. איכות התוצאות יחסית לגודל קבוצת המבחן הוא טוב ועומד על 96.75 (Accuracy). נתון זה הוא השני באיכותו מבין כלל שיטות הסיווג שנבחנו.

כאשר משווים בין הרשת הגנרטיבית לרשת קונבולוציה משולבת מנגנון הקשב, ניתן לראות שמדדי הדיוק כמעט זהים (הפרש של רבע האחוז) אך קבוצת המדגם של הרשת קונבולוציה-קשב גדולה פי כמה ובעלת כמות מחלקות גדולה יותר. וכאשר משווים בין רשת הקונבולוציה המשולבת עם תא LSTM לרשת הקונבולוציה המשולבת עם מנגנון הקשב ישנה עדיפות באחוזי הדיוק לזו הראשונה, אך גודל קבוצת המדגם של השנייה גדול בסדרי גודל משמעותיים.

אם כן קשה לתת את הדעת בצורה מפורשת כאשר מנסים להשוואת את איכות המודלים הנבחנו וזאת בשל הפערים בגודל קבוצת המבחן וכמות המחלקות המסווגות.

יחד עם זאת, מנקודת מבטי, הרשת המבוססת על מנגנון הקונבולוציה עם שכבת הקשב מציגה תוצאות טובות יחסית לרוב הרשתות האחרות, והיא בעלת קבוצת המדגם הגדולה ביותר בפער גדול. אי לכך ועל אף שהנתונים לא חד ערכיים, מנקודת מבטי היא משיגה את התוצאות הטובות ביותר. כמובן שכאשר בוחנים את איכות מדדי הערכה מבלי להתחשב בגודל קבוצת המדגם וכמות המחלקות המסווגות, רשת הקונבולוציה המשולבת עם שכבת LSTM מגיע לאחוזי הדיוק הטובים ביותר.

7. סיכום ומסקנות

עבודת גמר זו התמקדה בניתוח השוואתי של שיטות שונות מבוססות למידה עמוקה לסיווג צילומי רנטגן של בית החזה במטרה לזהות ולסווג תמונות של אנשים החולים במחלות ריאה שונות.

במהלך העבודה נסקר עולם הלמידה העמוקה, הוצגו מושגים בתחום רשת הנוירונים ונסקרו כלי האימון של הרשת. בנוסף נסקרו טכניקות שונות לבניית רשתות נוירונים בדגש על רשתות נוירונים קונבולוציוניות, רשתות נוירונים נשנות, מנגנון הקשב ורשתות נוירונים גנרטיביות.

לאחר מכן, הוצגו ארבע גישות שונות לביצוע סיווג של צילומי רנטגן באמצעות כלים שונים של רשתות נוירונים מתקדמות. השיטות כללו סיווג באמצעות רשתות קונבולוציה בלבד, סיווג באמצעות רשתות קונבולוציה המחוברות לתאי זיכרון נישנים, רשתות קונבולוציה המחוברות הקשב ורשתות קונבולוציה המחוברות לרשתות גנרטיביות.

על אף שמדדי ההערכה שצוינו בעבודה, להערכת התוצאות של הרשתות השונות מראים שאחוזי הדיוק בקרב חלק מהרשתות גבוה מאחרות, קיים קושי להשוות את ביצועי הרשתות וזאת מכיוון שגודל מאגרי הנתונים בין המחקרים השונים שהוצגו וכמות מחלקות הסיווג עבור כל מחקר איננו זהה. בנוסף, בחלק מהמחקרים הושם דגש על מדדי הערכה שונים בעולם ביצועי התוכנה – קרי זמן העיבוד וגודל המודל.

יחד עם זאת ניתן להגיד למנגנון בזהירות כי רשת הנוירונים המבוססת על שילוב של רשת קונבולוציה עם מנגנון קשב הציגה ביצועיים טובים יותר יחסית לגודל קבוצת המדגם וכמות המחלקות שסווגו במחקר.

בנימה אישית, העבודה תרמה לי מאוד בהבנת עולם הראייה הממוחשבת וכלי הלמידה העמוקה הנמצאים בשימוש בתחום זה.

8. המלצות למחקר עתידי

ניתן לתת מספר המלצות למחקר עתידי בתחום סיווג תמונות הרנטגן של בית החזה:

1. הרחבת מרחב הסיווג למחלות נוספות – תחום המחקר יכול לכלול לטעמי מחקר אחר מחלות ריאתיות נוספות. ככלל, מחלות אחרות אשר פוגעות ברמת חייהם של מטופלים ובאות לידי ביטוי בצילומי רנטגן של בית החזה יכולות להיות מאובחנות על ידי כלי ראייה ממוחשבת. זיהוי וסיווג מחלות נוספות יכול לתת כלים נוספים בקרב הקהילה הרפואית לטובת שיפור היכולת לאבחן ולזהות מחלות.
2. שיפור ביצועים – מלבד השיפור באחוזי הדיוק של המודלים השונים ניתן אף לשפר ולייעל את ביצועי המערכות מבחינת גודל מודל רשת הנוירונים וזמן הריצה שלו. ככול שהמחקר יתמקד במציאת מודלים קטנים יותר ומהירים יותר כך השימוש בכלי למידה עמוקה יגבר. זמן הריצה בזמן האימון ובמהלך אבחון יכול להיות משמעותי ולפגוע ביעילות המערכת. בנוסף, ככול שהמודלים יהיו קטנים יותר כך החומרה שתפעיל אותם תוכל להיות פחות מורכבת ויקרה.
3. חקר ביצועים ושילוב גישות מחקר – בעבודה זו הצגתי מספר רשתות נוירונים שהביאו לביצועים טובים מאוד במשימת הסיווג. יחד עם זאת לא הצגתי רשת נוירונים המשלבת את כלל סוגי הרשתות שנסקרו – קרי מודל בעל רשת גנרטיבית, הפועל לצד ובמקביל לרשת קונבולוציה משולבת שכבות קשב. שילוב האמצעים הללו יכול להביא למיצוי תהליך הסיווג.

לסיכום, מחקר עתידי יכול להתמקד בטיוב ושיפור תוצאות רשת הנוירונים לצד הגדלת מישור הבעיה הרפואית ואבחון מחלות ריאה נוספות לצד צמצום המשאבים הדרושים לאימון והפעלת המודלים בזמן הכנת המודלים והשימוש בהם לסיווג תמונות.

9. ביבליוגרפיה

1. Abhijit, B., Divyanshu, B., Sunil, K., Prayas, T., Rahul, S., & Ram, B. P. (2021). A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images. *Biomedical Signal Processing and Control*.
2. Abobaker, M., Qasem, F., & Shangming, Y. (2023, march 22). *Automatic lung disease classification from the chest X-ray images using hybrid deep learning algorithm*. Retrieved from <https://link.springer.com/article/10.1007/s11042-023-15047-z>
3. Alex, S. (2020, march). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Special Issue on Machine Learning and Dynamical Systems*.
4. Alexander, A. (2024). *Introduction to deep learning* . MIT EECS.
5. Aras, M. I., & Abdulkadir, s. (2020). Deep learning approaches for COVID-19 detection based on chest . *Elsevier Ltd*.
6. Asher, C. (2023). Understanding Neural Network Training: Forward Propagation and Backpropagation. *medium*.
7. Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., . . . Illia, P. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems* 30.
8. Aston, Z., Zack, C. L., Mu, L., & Alex, J. S. (2023). *Dive into Deep Learning*. Cambridge: Cambridge University Press.
9. Bat sheva, G. (2023, 11 26). *Pneumonia*. Retrieved from Clalit: [https://www.clalit.co.il/he/your_health/family/Pages/pneumonia.aspx#:~:text=%D7%A8%D7%99%D7%90%D7%95%D7%AA%20\(Pneumonia\)%3F-,%D7%93%D7%9C%D7%A7%D7%AA%20%D7%A8%D7%99%D7%90%D7%95%D7%AA%20\(%D7%91%D7%90%D7%A0%D7%92%D7%9C%D7%99%D7%AA%3A%20Pneumonia%2C%20%D7%A4](https://www.clalit.co.il/he/your_health/family/Pages/pneumonia.aspx#:~:text=%D7%A8%D7%99%D7%90%D7%95%D7%AA%20(Pneumonia)%3F-,%D7%93%D7%9C%D7%A7%D7%AA%20%D7%A8%D7%99%D7%90%D7%95%D7%AA%20(%D7%91%D7%90%D7%A0%D7%92%D7%9C%D7%99%D7%AA%3A%20Pneumonia%2C%20%D7%A4)

10. Dina, M. I., Nada, M. E., & Amany, M. S. (2021, may). *Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases*. Retrieved from Computers in Biology and Medicine:
<https://www.sciencedirect.com/science/article/pii/S0010482521001426>
11. Fairouz, H., Ala, M., Shadi, A., Subhieh, M. E.-S., Belal, A., Laith, A., & Amir, H. G. (2022, September 27). *Hybrid CLAHE-CNN Deep Neural Networks for Classifying Lung Diseases from X-ray Acquisitions*. Retrieved from Electronics:
<https://www.mdpi.com/2079-9292/11/19/3075>
12. Geng, H., Xiaoyan, C., Jianyong, C., Miao, Z., Yumeng, R., & Xinyu, Z. (2021). A multi-scale gated multi-head attention depthwise separable CNN model for recognizing COVID-19. *nature protfolio*.
13. Goram, M. M., Qiang, N., Richard, J., Haris, P., & Nada, M. E. (2023, February). A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. *Alexandria Engineering Journal*, pp. 923-935.
Retrieved from Alexandria Engineering Journal.
14. Ian, G., Yoshua, B., & Aaron, C. (2016). *Deep Learning*. MIT Press.
15. Ian, J. G., Jean, P.-A., Mehdi, M., Bing, X., David, W.-F., Sherjil, O., . . . Yoshua, B. (2014). Generative Adversarial Networks. *Universite de Montreal*.
16. Idan, a. (2023). *'Deep learning - course manuel* .The Open University.
17. islam, z., islam, m., & asraf, a. (2020). a combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using x-ray images. *khulna university of engineering and technology*.
18. Isoon, K., Kasi, T., Boonchana, P., & Attasit, L. (2023, january 25). *CNN–RNN Network Integration for the Diagnosis of COVID-19 Using Chest X-ray and CT Images*. Retrieved from Sensors: <https://www.mdpi.com/1424-8220/23/3/1356>

19. Jaleel, A. (2023). Gradient Descent From Scratch- Batch Gradient Descent, Stochastic Gradient Descent, and Mini-Batch Gradient Descent. *medium*.
20. Kaiming, H., Xiangyu, Z., Shaoqing, R., & Jian, S. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
21. Karan, S., & Jinan, F. (2020, april). *Capsule Networks: An Alternative Approach to Image Classification Using Convolutional Neural Networks*. Retrieved from ResearchGate:
https://www.researchgate.net/publication/340611859_Capsule_Networks_An_Alternative_Approach_to_Image_Classification_Using_Convolutional_Neural_Networks
22. Keiron, O., & Ryan, N. (2015). An Introduction to Convolutional Neural Networks. *Department of Computer Science, Aberystwyth University, Ceredigion*.
23. Nick, V. (2022, 6). *Lung Opacity: What You Should Know*. Retrieved from healthline: <https://www.healthline.com/health/lung-opacity>
24. Sepp, H., & Jürgen, S. (1997). Long Short-term Memory. *Neural Computation*.
25. Shimpy, G., & Rajiv, S. (2021, september 18). *Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques*. Retrieved from Journal of Ambient Intelligence and Humanized Computing: <https://link.springer.com/article/10.1007/s12652-021-03464-7>
26. So-Mi, C., Seung-Seok, L., & Bonggyun, K. (2021, January 29). *Attention-Based Transfer Learning for Efficient Pneumonia Detection in Chest X-ray Images*. Retrieved from <https://www.mdpi.com/2076-3417/11/3/1242>
27. Steven, A. H., Inga, S., Vajira, T., Malek, H., Michael, A. R., Pål, H., & Sravanthi, P. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*.

28. Talya, K. (2013, APR). *Medical guides - chest x-ray*. Retrieved from clalit:
https://www.clalit.co.il/he/medical/medical_checkups/Pages/chest_x-ray.aspx
29. world health organization. (2023, mar). *coronavirus disease*. Retrieved from world health organization:
<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>
30. Yaman, A. (2023, Dec 2). *Automated Pneumonia Based Lung Diseases Classification with Robust Technique Based on a Customized Deep Learning Approach*. Retrieved from Department of Software Engineering, Faculty of Technology, Firat University, Elazig 23200, Turkey:
<https://www.mdpi.com/2075-4418/13/2/260>
31. Zabirul, I., & Milon, I. (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Department of Computer Science and Engineering, Khulna University of Engineering & Technology*.

Abstract

Many diseases harm the quality of life of humans and can cause severe symptoms and sometimes death. Lung diseases can be diagnosed in diverse ways, among them a chest x-ray. X-ray is a central, available, cheap, and convenient tool for diagnosing Covid19, viral pneumonia, lung cancer and other diseases. The diagnosis is made by specialist doctors who can exam the X-ray image of the chest, which contains patterns that indicate lung disease.

The process of diagnosis and classification is manual, long and complex which requires medical expertise and sometimes cannot be conducted in real time due to the absence of a doctor with appropriate expertise. Therefore, an automatic system for classifying lung diseases based on a chest X-ray is required that can classify the photograph into the correct type of disease.

Such a system will allow appropriate treatment. During the last few years, in light of the development of computational capabilities and the growth of the field of deep learning the use of deep learning for the purposes of identifying and classifying X-rays has increased.

In this work, a number of methods from the field of deep learning were reviewed for classifying chest X-rays for healthy people, people with covid19 and people with other lung diseases.

Special emphasis is given in the review to neural networks based on convolution alongside their combination with adaptive neural networks, attention layers and generative neural networks.

In the work, a comparative analysis was performed between the variety of deep learning methods, which was based on various quality indicators such as accuracy, sensitivity, and uniqueness.

In this comparative analysis, it was found that there are several effective methods for classifying lung diseases from X-rays, and in particular it was found that the CNN-ATTENTION type model produces accurate results relative to the size of the sample group examined and relative to the number of departments classified in the study.

Contents

1.	Abstract.....	2
2.	introduction	7
2.1	Work structure description.....	7
2.2	Medical problem definition.....	7
2.2.1	Diagnosis of lung diseases	7
2.2.2	COVID-19 and its effects	7
2.2.3	Pneumonia.....	8
2.2.4	Lung opacity	9
2.3	Purpose and importance	10
3.	Deep learning as a classification tool.....	12
3.1	Deep neural networks	12
3.1.1	Setting up a deep neural network	12
3.1.2	The learning process	14
3.1.3	Network authentication	21
3.1.4	Examining the network	22
3.1.5	Overfitting	22
3.1.6	Tools to prevent overfit and increase generalization	23
3.2	Convolution networks.....	25
3.2.1	Typical layers of convolutional networks.....	26
3.2.2	Authentication training and network test	31
3.2.3	Widely used convolution networks.....	32
3.3	Recurring networks (RNN)	34
3.3.1	Network training process	36
3.3.2	Authentication and recurrent network test.....	37
3.3.3	Recurring network architecture	38
3.3.4	The problem of the disappearing and exploding gradient ...	38

3.3.5	LSTM	39
3.4	ATTENTION mechanism and Transformers	41
3.4.1	Schematic description of ATTENTION	41
3.4.2	Formal description of of ATTENTION.....	42
3.4.3	Transformers.....	44
3.5	Generative networks	48
3.5.1	Generative network structure	48
3.5.2	Network Training	49
3.5.3	The loss function.....	51
3.5.4	Test and performance evaluation in generative networks ...	53
3.5.5	Problems with the loss function in generative networks.....	54
4.	Methods for classifying pulmonary diseases using X-rays	56
4.1	Classification of pulmonary diseases using convolution.....	56
4.1.1	A Case Study for Classifying Lung Diseases Using Convolution Networks	58
4.1.2	Database	58
4.1.3	Research Method	59
4.1.4	Study Results and Conclusions	61
4.2	Classification of pulmonary diseases using CNN-RNN	62
4.2.1	Case study for classification using LSTM-CNN	64
4.2.2	Database	64
4.2.3	Research Method	64
4.2.4	Study Results and Conclusions	67
4.3	Classification of pulmonary diseases using CNN- ATTENTION ..	69
4.3.1	Case Study for Classification with Convolution Network Integrated Attention Mechanism	70
4.3.2	Database	71
4.3.3	Study Course	71

4.3.4	Study Results and Conclusions	73
4.4	Classification of pulmonary diseases using GAN-CNN	76
4.4.1	introdaction	76
4.4.2	Study Objectives	76
4.4.3	Database	77
4.4.4	Study Course	77
4.4.5	Study Results and Conclusions	82
5.	Evaluation metrics for comparative analysis	84
5.1	Accuracy	85
5.2	Precision	86
5.3	Sensitivity	87
5.4	Specificity	87
<u>5.5</u>	ROCK Curve	88
5.6	Loss function	89
6.	Comparative analysis of the presented methods	91
7.	Conclusions	94
8.	Recommendations for future research	95
9.	Bibliography	96

The Open University of Israel

Department of Mathematics and Computer Science

A Comparative Review of Methods for Classifying Lung Diseases from Radiographs

Final paper submitted as partial fulfillment of the requirements

Towards an M.Sc. degree in Computer Science

The Open University of Israel

Department of Mathematics and Computer Science

BY

Tal Kfir

Prepared under the supervision of Dr. Maya Herman

MAY 2024