# פרויקט למידת מכונה

בחינת ביצועי רשתות קונבולוציוניות ומכונות וקטור תמיכה על מאגר MNIST

**מוגש למרצה: דוקטור טום טריגנו**

**סטודנטים:**

**אביתר חיימוף – 208648006**

**ניר בנימין - 213922487**

תוכן עניינים

[פרויקט למידת מכונה 1](#_Toc196650724)

[תוכן עניינים 1](#_Toc196650725)

[מבוא 2](#_Toc196650726)

[מודלים 2](#_Toc196650727)

[SVM 3](#_Toc196650728)

[Dataset 3](#_Toc196650729)

[קריטריוני טיב: 4](#_Toc196650730)

[מהלך העבודה 5](#_Toc196650731)

[הכנת הנתונים: 5](#_Toc196650732)

[בניית מודלי SVM 5](#_Toc196650733)

[פיתוח רשת CNN 6](#_Toc196650734)

[שמירת התוצאות והצגתן 6](#_Toc196650735)

[התפלגות הנתונים עבור CNN 6](#_Toc196650736)

[התפלגות הנתונים עבור SVM 7](#_Toc196650737)

[תוצאות ומסקנות 8](#_Toc196650738)

[תוצאות מודל SVM 8](#_Toc196650739)

[תוצאות מודל CNN: 10](#_Toc196650740)

[נשווה כעת בין המודלים ונבחר את הטוב ביניהם 12](#_Toc196650741)

[נספחים 13](#_Toc196650742)

# מבוא

העבודה עוסקת בבדיקת ביצועי שתי שיטות סיווג שונות על מאגר הנתונים MNIST, המשמש לזיהוי ספרות בכתב יד. בשיטה הראשונה נבנה רשת עצבית קונבולוציונית (CNN), המייצגת גישה מודרנית של למידה עמוקה לעיבוד תמונה, ובשיטה השנייה נבחר להשתמש במכונת וקטור תמיכה (SVM), אשר מהווה גישה קלאסית יותר. העבודה כוללת שלבים של טעינת הנתונים, עיבוד מוקדם, בניית המודלים, אימון ובדיקת ביצועי כל שיטה לפי מדדים כגון דיוק וזמן אימון. באמצעות השוואה זו, נבחן את היתרונות והחסרונות של כל שיטה ונציג תובנות שיכולות לסייע בבחירת גישה מתאימה לבעיות סיווג מורכבות יותר.

## מודלים

**CNN**

בכניסה לרשת מתקבלת תמונה בגודל 28×28 של ספרה בכתב יד מתוך מאגר הנתונים MNIST.

הרשת כוללת שתי שכבות קונבולוציה:

* השכבה הראשונה מפיקה 32 מפות תכונה (feature maps) באמצעות גרעין (kernel) בגודל 3×3.
* השכבה השנייה מפיקה 64 מפות תכונה, גם היא באמצעות גרעין בגודל 3×3.
* לאחר כל שכבת קונבולוציה מופעלת פונקציית האקטיבציה ReLU.
* בתום כל שלב קונבולוציה מתבצעת פעולת מקסימום בריכה (MaxPooling) בגודל 2×2 ו־Stride של 2, לצורך הקטנת המימדים.

בשל תהליך הקונבולוציה וה-Pooling, מימדי מפת הפיצ'רים הסופית הם 7×7 עם 64 ערוצים, ולכן הקלט לשכבה הצפופה הראשונה הוא בגודל 64·7·7 = 3136 מאפיינים.

הרשת כוללת שתי שכבות Fully Connected:

* השכבה הראשונה מכילה 128 נוירונים.
* השכבה השנייה מכילה 10 נוירונים, אשר מייצגים את 10 הספרות האפשריות (0–9). על פלט שכבה זו מופעלת פונקציית Softmax לצורך חישוב הסתברויות.

בנוסף, מיושמת רגולריזציה מסוג Dropout בשיעור של 25% (0.25) להפחתת Overfitting, ופונקציית המחיר בה נעשה שימוש היא Cross-Entropy.

האימון:

נאמן את המודל במשך 100 Epochs.

Early Stopping: נפסיק את האימון של המודל לאחר חוסר שיפור ה-Validation Loss לאחר 5 Epochs.

מייעל ADAM לקצב למידה משתנה לעדכון המשקולות.

## SVM

ניישם גם מודל **SVM (Support Vector Machine)** לצורך השוואת ביצועים.  
נבחן שלושה גרעינים שונים:

* **לינארי (Linear)**
* **גאוסי (RBF – Radial Basis Function)**
* **פולינומיאלי (Polynomial)**

לצורך בחירת הפרמטרים האופטימליים, נשתמש ב־**Grid Search בשילוב הצלבה של 5-fold Cross-Validation**. החיפוש יתבצע על פני הטווחים הבאים:

* C: [0.1, 1, 10]
* gamma: [0.01, 0.05, 0.1]
* kernel: ['linear', 'rbf', 'poly']
* degree: [2, 3, 4]

בכל שילוב פרמטרים ייאמד הביצוע באמצעות דיוק על סט אימות. שילוב הפרמטרים שיניב את התוצאה הטובה ביותר ייבחר, וביצועיו יושוו לתוצאות שהתקבלו באמצעות רשת ה־CNN – הן מבחינת דיוק, והן מבחינת זמן הרצה ויכולת הכללה. אחנחנו נבחר את הפרמטרים שמחזיר ה -Grid Search.

## Dataset

בפרויקט נעשה שימוש במאגר הנתונים **MNIST** (Modified National Institute of Standards and Technology), שהוא אחד ממאגרי הנתונים הנפוצים ביותר בתחום של זיהוי ספרות בכתב יד. המאגר כולל **70,000 תמונות בגווני אפור בגודל 28×28 פיקסלים**, מהן **60,000 משמשות כקבוצת אימון** ו־**10,000 כקבוצת בדיקה**. כל תמונה מייצגת ספרה אחת (0–9) שנכתבה בכתב יד על ידי מגוון כותבים, מה שמקנה למאגר גיוון רב באופני הכתיבה.

תמונות המאגר עברו תהליך עיבוד מקדים, הכולל חיתוך, סידור במרכז התמונה ונירמול הגודל, כך שהספרה מופיעה במרכז התמונה ומוקפת ברקע שחור. תכונות אלו הופכות את MNIST לבסיס נתונים אידיאלי ללמידת מכונה ולניסויים בתחום של ראייה ממוחשבת וסיווג תמונות.

קישור: <https://www.kaggle.com/datasets/hojjatk/mnist-dataset>

## קריטריוני טיב

להערכת ביצועי המודלים בפרויקט נשתמש בארבעה קריטריוני טיב מרכזיים:

1. Accuracy - אחוז הדוגמאות שסווגו נכון מתוך כלל הדוגמאות. מדד כולל שמתאר את ביצועי המודל באופן כללי.
2. Confusion Matrix - מטריצת בלבול) – מטריצה בגודל 10×10 אשר מציגה את התפלגות התחזיות מול התשובות האמיתיות לכל אחת מהספרות, ומאפשרת לזהות דפוסי טעות בין מחלקות.
3. Precision - עבור כל מחלקה, נבחן את אחוז הדוגמאות שסווגו למחלקה מסוימת והיו נכונות בפועל. מדד זה חשוב במיוחד כאשר חשוב לנו לא "להאשים" ספרה שגויה בטעות.
4. Recall - עבור כל מחלקה, נבחן את אחוז הדוגמאות מהמחלקה שסווגו נכון. מדד זה חשוב כאשר אנו רוצים לוודא שכל המחלקות מזוהות כראוי.

המדדים יחושבו עבור כל מחלקה בנפרד וכן באמצעות ממוצע משוקלל (weighted average) לצורך השוואה כללית בין המודלים.

**מאמר** עליו אנו מתבססים: "A Survey of Handwritten Character Recognition with MNIST and EMNIST"

קישור: <https://www.mdpi.com/2076-3417/9/15/3169>

חשוב לציין כי מאמר זה משווה בין שיטות סיווג, ולא מציין כיצד לממש אותן.

# מהלך העבודה

## הכנת הנתונים:

החלק החשוב ביותר כאשר עובדים עם נתונים הוא להכיר אותם, ללמוד את תכונותיהם ובהתאם לכך להתאים אותם לצורכי העבודה.

בתחילת התהליך טענו את מאגר הנתונים MNIST, המורכב מ־60,000 תמונות לסט האימון ו־10,000 תמונות לסט הבדיקה.  
לאחר טעינת הנתונים, ביצענו נירמול בעזרת StandardScaler, אשר קבע את טווח הערכים לכל התמונות בין ‎0 ל־1.  
נירמול זה אפשר השוואה הוגנת יותר בין המודלים תוך שמירה על עקביות בנתונים.

## בניית מודלי SVM

לצורך אימון מודלי ה־SVM, צמצמנו את סט האימון ל־10,000 תמונות בלבד, עקב מגבלות זיכרון וזמן חישוב.

בנינו מודלים עם שלושה סוגי גרעינים:

* לינארי (Linear)
* גאוסי (RBF)
* פולינומיאלי (Polynomial)

לבחירת הפרמטרים האופטימליים בוצע Grid Search בשילוב חמישה־קיפולים (5-Fold Cross Validation), תוך חיפוש בערכים שונים של C, gamma ו־degree.

## פיתוח רשת CNN

בנינו רשת קונבולוציונית בסיסית אך חזקה, הכוללת שתי שכבות קונבולוציה ואחריהן שכבות Fully Connected.

למניעת Overfitting יושמו טכניקות רגולריזציה:

* Dropout בשיעור של 25%.
* Early Stopping כאשר לא נצפה שיפור נוסף ב־Validation Loss.

## שמירת התוצאות והצגתן

עבור מודלי ה־SVM שמרנו את המודלים שהשיגו את הביצועים הטובים ביותר, והצגנו:

גרפים של תוצאות ה־Grid Search

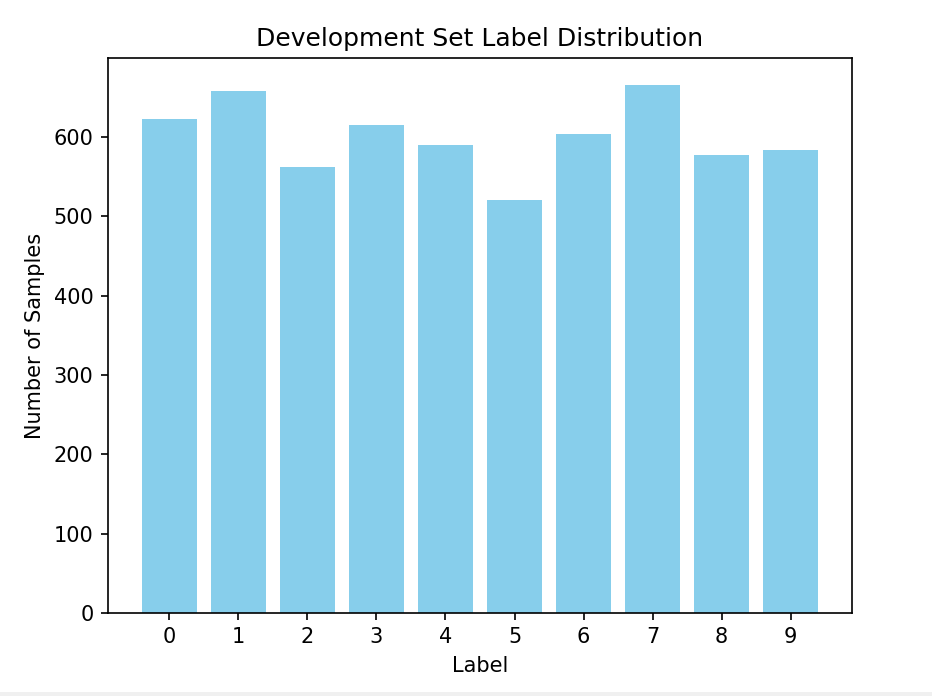
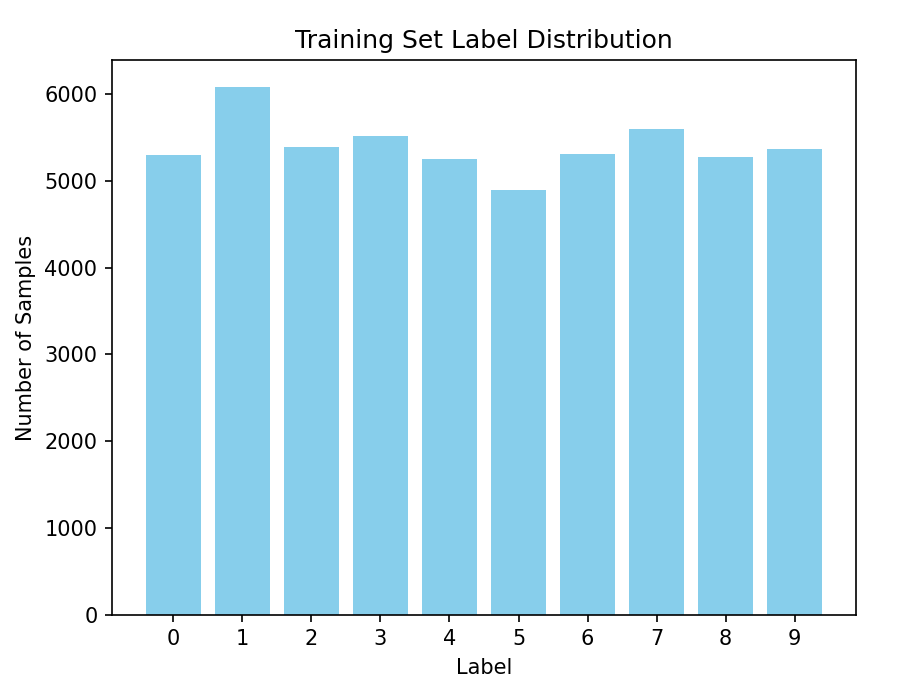
מטריצות בלבול לכל גרעין

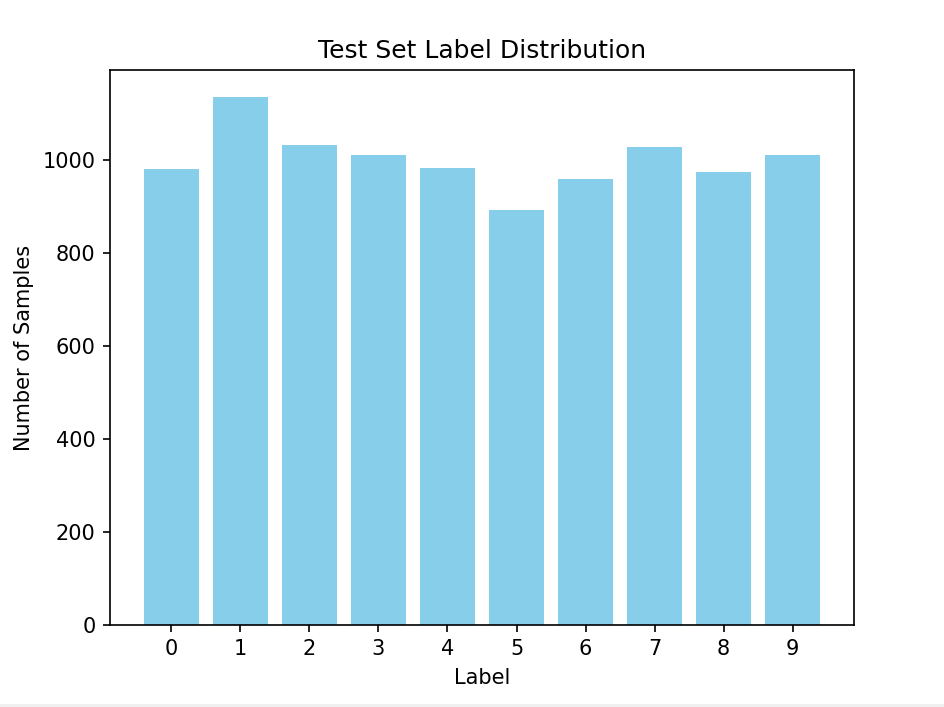
עבור מודל ה־CNN שמרנו את המודל בעל התוצאה הגבוהה ביותר, והצגנו:

עקומות למידה של Loss ו־Accuracy

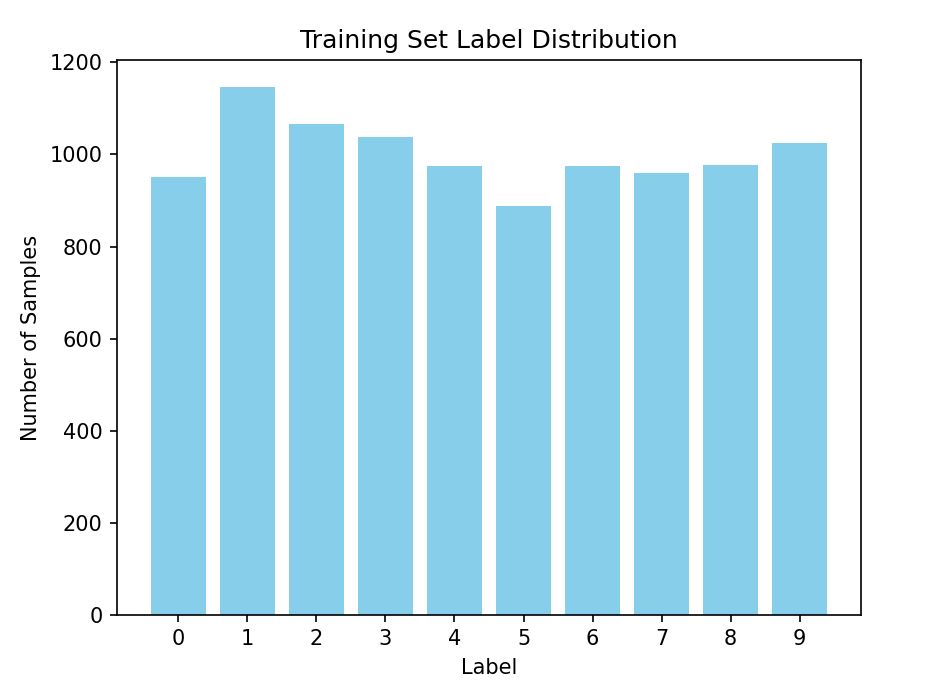
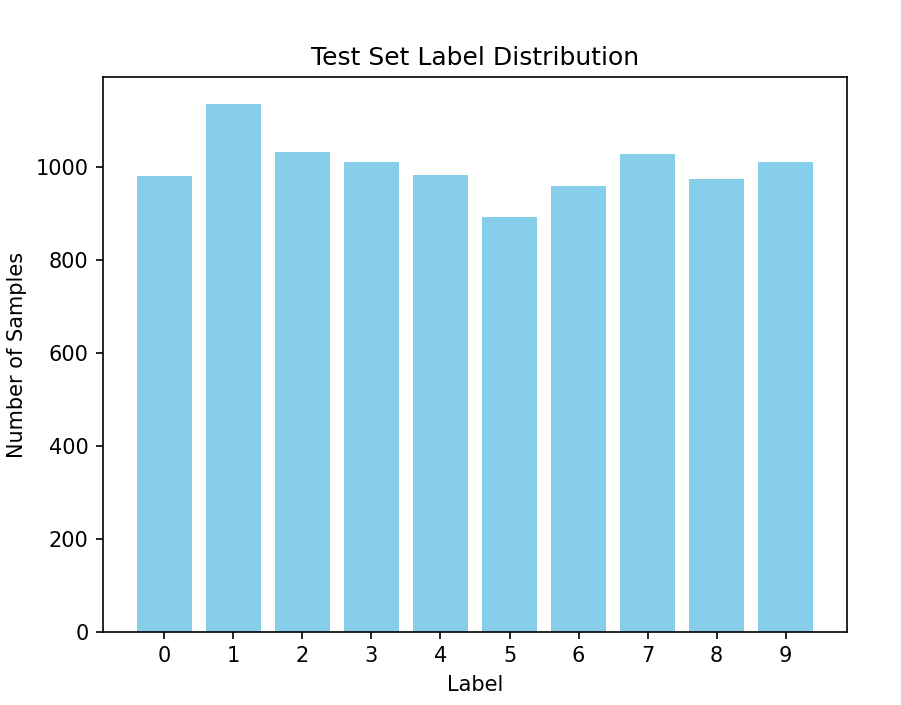
מטריצת בלבול מסכמת.

## התפלגות הנתונים עבור CNN





## התפלגות הנתונים עבור SVM





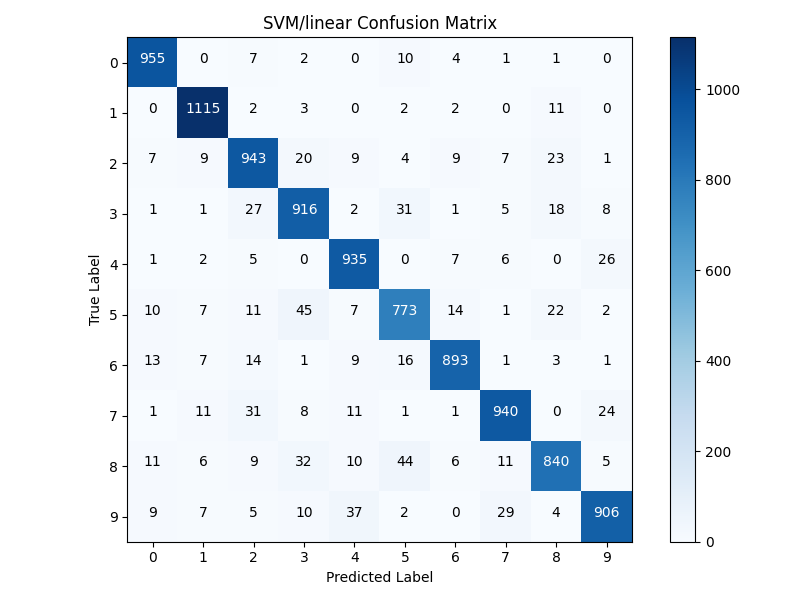
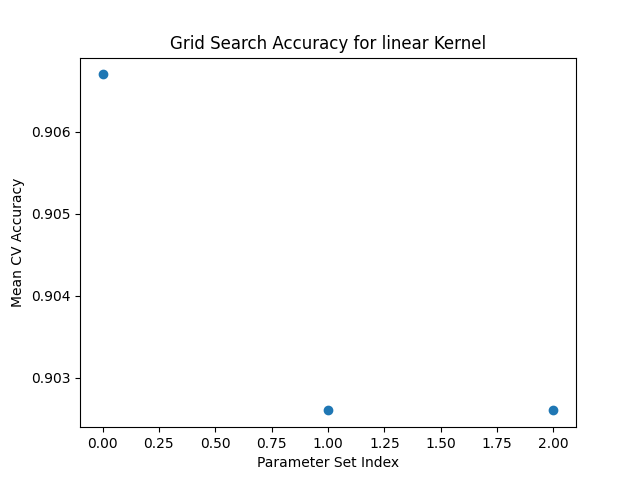
כפי שציינו קודם לכן, סט האימון עבור מודל זה כולל 10,000 תמונות.

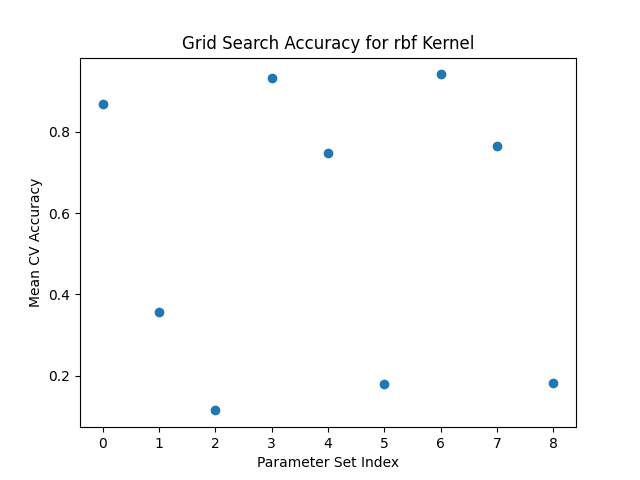
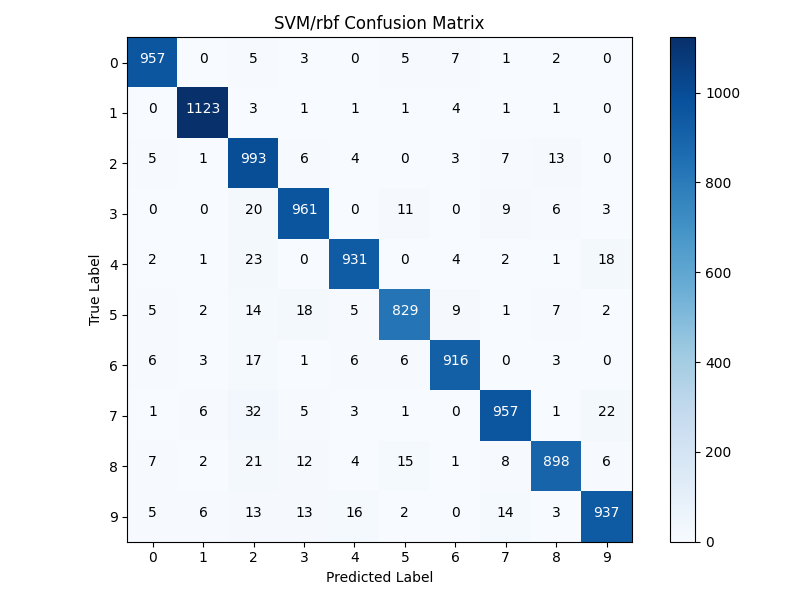
במודל SVM אין צורך להפריד את הנתונים לסט ולידציה, כפי שעשינו באימון רשת ה־CNN.  
הסיבה לכך היא שמודל SVM מבוסס על פתרון בעיית אופטימיזציה מוגדרת מראש, שמטרתה מציאת ההפרדה הטובה ביותר בין הדגימות.

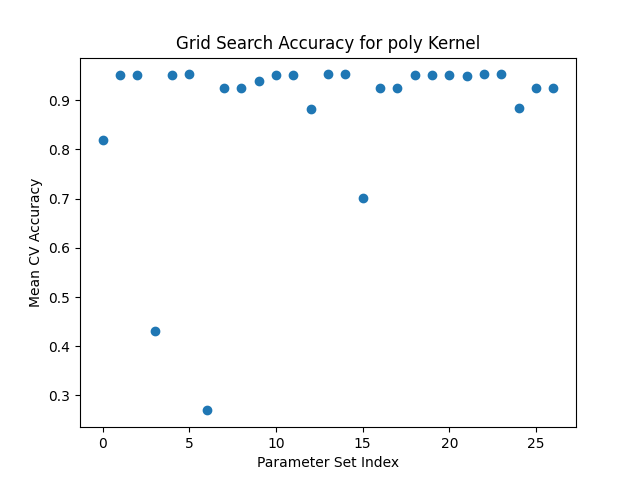
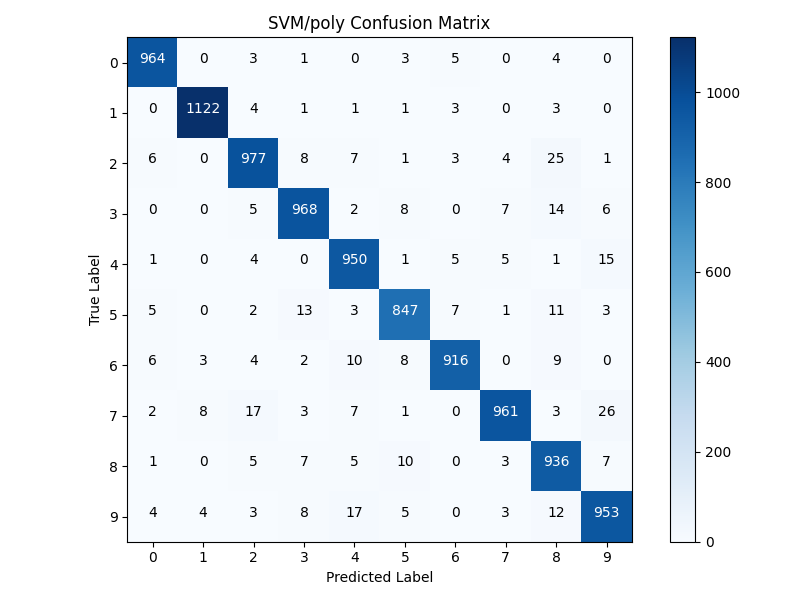
בנוסף, במהלך תהליך ה־Grid Search, מתבצע Cross Validation באופן אוטומטי: הנתונים מחולקים לתת־קבוצות, כך שמתאפשרת הערכת ביצועים אמינה תוך כדי מניעת תופעות כמו Overfitting.

# תוצאות ומסקנות

## תוצאות מודל SVM

**Linear kernel**

**RBF kernel**

**Polynomial kernel:**

**נסכם את הביצועים של הגרעינים בטבלה:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kernel** | **Accuracy** | **precision** | **recall** | **Test time (s)** |
| **Linear** | 92.16% | 92.17% | 92.16% | 19.26 |
| **RBF** | 95.94% | 95.97% | 95.94% | 15.38 |
| **Polynomial** | 95.02% | 95.11% | 95.02% | 23.33 |

כפי שניתן לראות מהטבלה, קיימים הבדלים בביצועים בין סוגי הגרעינים שנבחנו:

גרעין RBF השיג את התוצאות הגבוהות ביותר בכל המדדים בנוסף, זמן הבדיקה שלו היה הקצר ביותר (15.38 שניות), מה שמעיד על יעילות גבוהה גם בזמן ריצה.

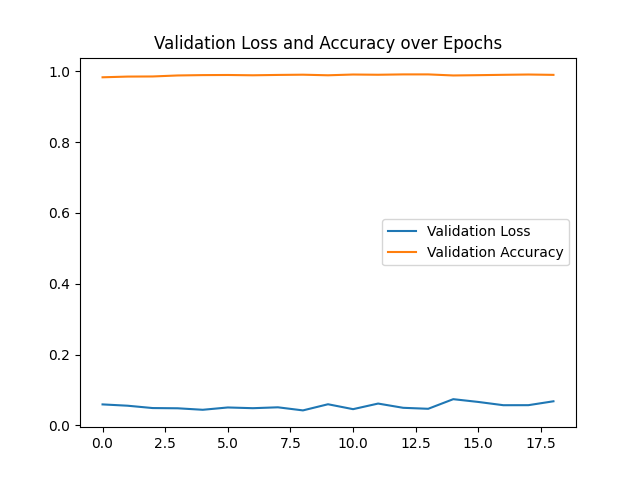
גרעין Polynomial הגיע לביצועים טובים מאוד, עם דיוק של 95.02%, אולם דרש זמן בדיקה ארוך יותר (23.33 שניות).

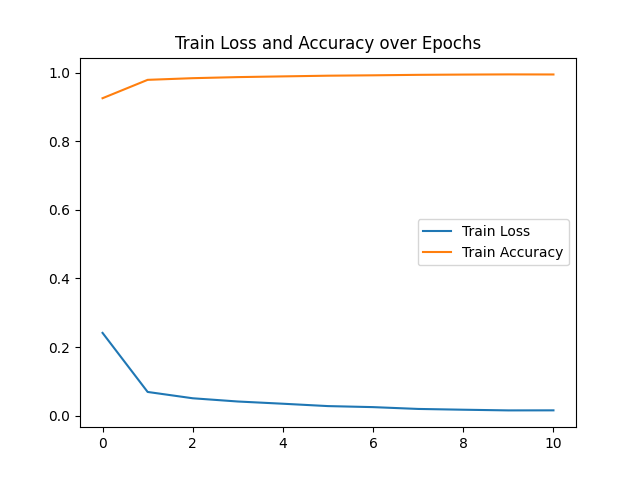
גרעין Linear, לעומת זאת, הראה את הביצועים הנמוכים ביותר, עם דיוק של 92.16%. יחד עם זאת, זמן הבדיקה שלו היה סביר (19.26 שניות), מה שיכול להפוך אותו לבחירה מתאימה כאשר פשטות המודל וזמן חישוב מהירים הם דרישות עיקריות.

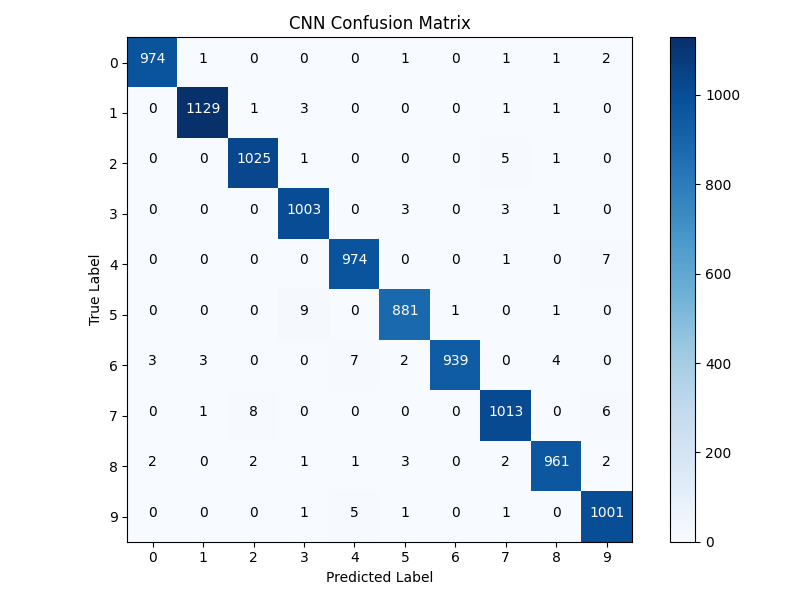
מסקנה ראשונית היא כי עבור סיווג ספרות ממאגר MNIST, שימוש בגרעין RBF במודל SVM מניב את התוצאות הטובות ביותר הן מבחינת איכות הסיווג והן מבחינת יעילות חישובית.

פרמטרי הגרעינים עבורם קיבלנו את התוצאות הטובות ביותר:

## תוצאות מודל CNN:

* עבור מודל סט האימון נשאר בגודלו המקורי





נסכם את התוצאות בטבלה:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **precision** | **recall** | **Test time (s)** |
| **CNN** | 99% | 99% | 99% | 0.30 |

כפי שניתן לראות, המודל הצליח להגיע לתוצאות טובות במיוחד. הוא הגיע לדיוק של 99%, עם תוצאות עקביות מאוד ב־precision ו־recall.

עקומות הלמידה מראות ירידה חדה בלוס ועלייה יציבה בדיוק, אשר מעידים על כך שאין סימנים משמעותיים ל overfitting. מטריצת הבלבול מצביעה על דיוק גבוה כמעט בכל המחלקות, עם טעויות בודדות בלבד.

## נשווה כעת בין המודלים ונבחר את הטוב ביניהם

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **precision** | **recall** | **Test time (s)** |
| **CNN** | 99% | 99% | 99% | 0.30 |
| **Linear** | 92.16% | 92.17% | 92.16% | 19.26 |
| **RBF** | 95.94% | 95.97% | 95.94% | 15.38 |
| **Polynomial** | 95.02% | 95.11% | 95.02% | 23.33 |

מהשוואת הביצועים עולה באופן ברור כי מודל ה־CNN מצטיין בכל המדדים: דיוק, Precision ו־Recall של 99%, וזמן ריצה קצר ביותר של 0.30 שניות בלבד.

בהשוואה לכך, מודלי ה־SVM (לינארי, פולינומיאלי ו־RBF) הגיעו לדיוקים נמוכים יותר, וזמני הריצה שלהם ארוכים משמעותית — במיוחד עבור הגרעין הפולינומיאלי.

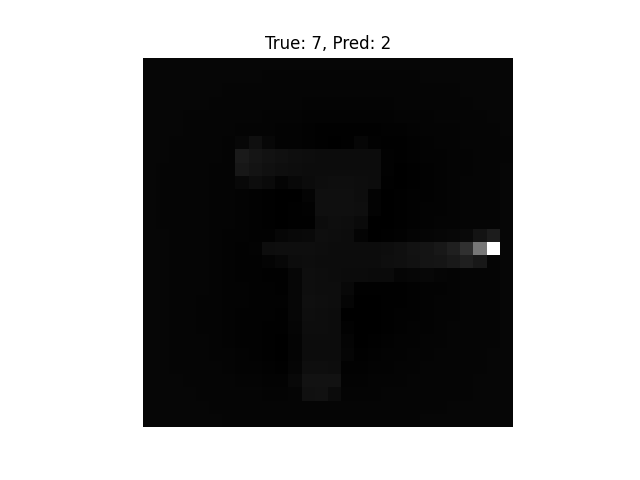
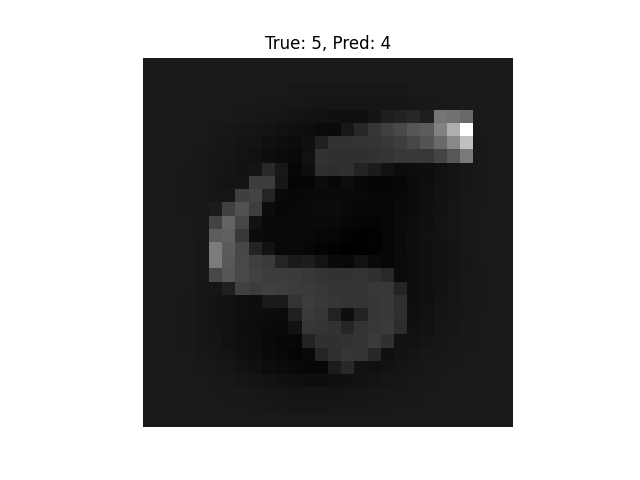
הגרעין שהראה את הביצועים הטובים ביותר מבין מודלי ה־SVM היה RBF, אך גם הוא נחות לעומת ה־CNN הן בדיוק והן ביעילות החישובית.

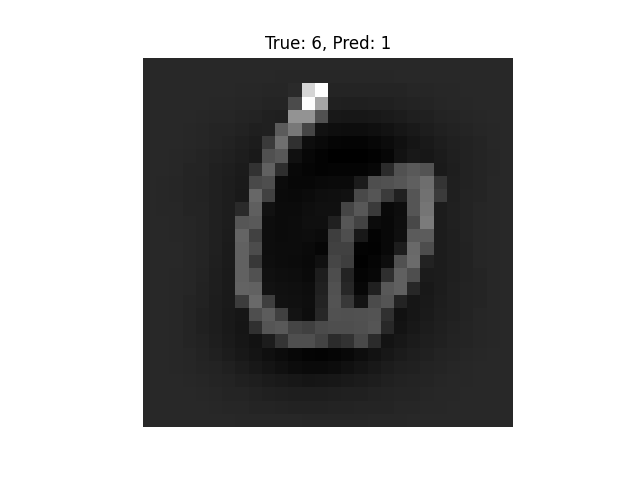
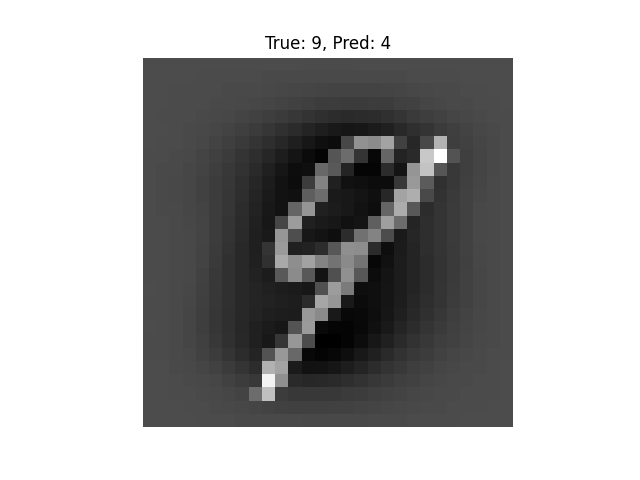
לכן, ניתן לקבוע בבירור כי מודל ה־CNN הוא הבחירה המועדפת למשימת סיווג ספרות בכתב יד ממאגר MNIST, כאשר קיימת גישה למשאבים המאפשרים אימון רשת נוירונים.

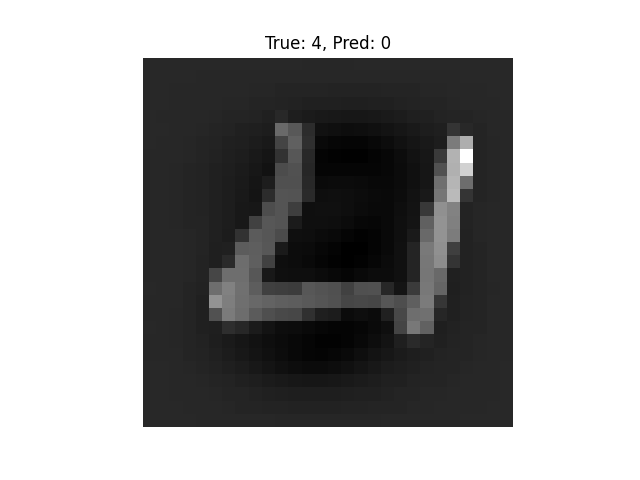
לסיכום בפרויקט זה, השווינו בין ביצועי מודלי CNN לבין מודלי SVM שונים על מאגר MNIST.

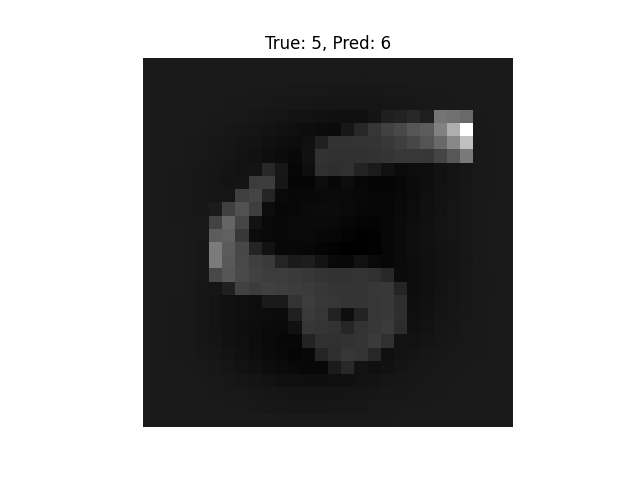
מצאנו כי מודל ה־CNN השיג תוצאות מעולות של 99% בכל מדדי הביצועים תוך זמן בדיקה קצר מאוד. בין מודלי ה־SVM, גרעין RBF הראה את התוצאות הטובות ביותר, אך עדיין נותר מאחור ביחס ל־CNN הן בדיוק והן ביעילות.

# נספחים

תוצאות שגויות עבור SVM (Poly)

תוצאות שגויות עבור SVM (Linear)

תוצאות שגויות עבור SVM (RBF)



תוצאות שגויות עבור CNN

