



המרכז האקדמי רופין
המחלקה למדעי המחשב

פיתוח מודל למידת מכונה לחיזוי משתני בית שורשים בחממה

דוח מסכם חלק א' זה הוגש כחלק מהדרישות לקבלת תואר

"בוגר במדעים" BSc במדעי המחשב

במרכז האקדמי רופין

המחלקה למדעי המחשב

על-ידי

עדן אליהו

אורן צ'אושו

העבודה הוכנה בהנחייתם של ד"ר אבנר פריאל וד"ר עלאא ג'מאל



תוכן עניינים:

1.	תקציר	4
1.1	מטרת העבודה	4
1.2	הערך היישומי	4
1.3	שיטת המחקר	4
1.4	תוצאות ראשוניות	5
2.	מבוא	
2.1	רקע כללי: האתגר בבקרת אקלים ובית שורשים	6
2.2	הפער בין הרצוי למצוי בניטור	6
2.3	הבסיס הפיזיקלי-ביולוגי	7
2.4	למה למידת מכונה?	8
2.5	הפתרון המוצע- חיישן וירטואלי	8
2.6	מטרת הפרויקט	8
3.	מושגים בסיסיים	
3.1	מושגים אגרונומיים	9
3.2	מושגים בלמידת מכונה	9
4.	סקירת ספרות	10
5.	הגדרת הבעיה	
5.1	ניסוח הבעיה	11
5.2	שאלת המחקר	11
5.3	השערת המחקר	11
6.	גישת הפתרון	
6.1	מסד הנתונים	12
6.1.1	נתונים רציפים	12
6.1.2	נתונים בדידים	12
6.2	השלמת נתונים היסטוריים	13
6.3	עיבוד מקדים וסנכרון זמנים	14
6.4	סביבת הפיתוח	14
6.5	ארכיטקטורת המערכת	14
6.5.1	מודל מיקרו-אקלים	14
6.5.2	מודל בית השורשים	14
6.6	אסטרטגיית איחוד הנתונים	15
6.7	הנדסת פיצ'רים	16
6.8	אסטרטגיית אימון ובדיקה	16
7.	תוצאות ראשוניות ואתגרים	
7.1	תוצאות המודלים	17
7.1.1	מודל המיקרו-אקלים	17
7.1.2	מודל בית השורשים	18
7.2	אתגרים ודרכי פתרון	19
7.2.1	היעדר ניטור רציף	19
7.2.2	בעיית הדלילות	19
7.2.3	רגישות לערכי EC נמוכים	19



20.....	7.2.4. אי סדירות זמנית של הדגימות
20.....	7.2.5. תלות במצב התחלתי ואפקט הזיכרון
21.....	7.2.6. אי רציפות בנתוני אקלים

	8. תוכנית עבודה ולוח זמנים
22.....	8.1. הנדסת פיצ'רים ושיפור המודל
22.....	8.2. אופטימיזציה של האלגוריתם
23.....	8.3. ניתוחי מעטפת וסימולציות
23.....	8.4. סיכום וכתובה

איורים:

Figure 1: Current Workflow – Manual Sampling with Late Detection and Reactive Correction

Figure 2: The Physical-Biological Basis – Factors influencing rootzone variables (EC & pH) resulting from the balance between Irrigation, Climate, and Plant Activity

Figure 3: Data Sparsity Problem – Continuous vs. Discrete Data

Figure 4: Historical Data Imputation – Results of the Microclimate Data Imputation Model

Figure 5: System Architecture – Two-Stage Prediction Architecture

Figure 6: Microclimate Model Results

Figure 7: Preliminary Results – Rootzone Model



1. תקציר:

1.1. מטרת העבודה :

מטרת פרויקט זה היא לפתח מודל חיזוי דינמי מבוסס למידת מכונה (Machine Learning) לניהול אופטימלי של בית השורשים בחממה.

הפרויקט מתמקד ביצירת מערכת המסוגלת לחזות את ערכי המוליכות החשמלית (EC) והחומציות (pH) העתידיים, בהינתן נקודת מוצא ידועה (מדידה נוכחית).

המודל משקלל שלושה וקטורים עיקריים כדי לבצע את החיזוי:

1. **מצב התחלתי:** נתוני ה-EC וה-pH שנמדדו בנקודת הזמן הנוכחית.
2. **תנאי סביבה:** תחזית מיקרו-אקלים בתוך החממה (קרינה, טמפרטורה, לחות והתאדות).
3. **ממשק גידול:** פרוטוקול ההשקיה והדישון המתוכנן (כמויות מים, הרכבי דשן ומועדי השקיה).

1.2. הערך היישומי:

יכולת חיזוי זו נועדה לשמש ככלי אופטימיזציה עבור המגדל. היא מאפשרת לעבור מגישה "תגובתית" (תיקון ערכים חריגים לאחר שכבר התרחשו) לגישה "פרואקטיבית" – זיהוי מגמות של המלחה או שינוי חומציות מראש, והתאמת משטר הדישון בזמן אמת כדי לשמור על תנאים אידיאליים לצמח, לחסוך במשאבים ולמקסם את היבול.

1.3. שיטת המחקר :

שיטת המחקר ומקורות הנתונים: המחקר מתבסס על אינטגרציה של נתונים מחמישה מקורות מידע שונים, היוצרים יחד תמונת מצב מלאה של סביבת הגידול:

1. **נתוני אקלים חיצוני:** נאספו מהשירות המטאורולוגי (תחנת בית דגן) וכוללים קרינה, טמפרטורה, לחות ומהירות רוח.
2. **נתוני חיישנים בחממה:** ניטור רציף של המיקרו-אקלים (אוויר) ונתוני קרקע שנאספו באמצעות Data Logger.
3. **נתוני תכנון גידול (ממשק):** פרוטוקולים מתוכננים של משטר ההשקיה והרכבי הדישון (כמויות ומועדים).
4. **נתוני התפתחות הצמח:** מעקב אחר כיסוי הנוף (Canopy Cover) המייצג את גודל הצמח וצריכת המים שלו.
5. **דגימות קרקע ידניות:** מדידות מעבדה תקופתיות ומדיקות של EC ו-pH המשמשות כעוגן לאימון המודל.



על בסיס נתונים אלו פותחה שרשרת מודלים (Pipeline) דו-שלבית:

מודל מיקרו-אקלים: חיזוי הטמפרטורה, הלחות וההתאדות העתידיים בתוך החממה על בסיס נתוני חוץ.

מודל בית שורשים: חיזוי EC ו pH על בסיס המיקרו-אקלים החזוי, היסטוריית ההשקיה והדישון, ומצב הצמח. נעשה שימוש באלגוריתמים מסוג LightGBM ו XGBoost- תוך שימוש בשיטת אימות Walk-Forward Validation המדמה תרחיש זמן-אמת.

1.4. תוצאות ראשוניות:

ממצאים ראשוניים: נכון לשלב זה של המחקר, המודל מציג תובנות מעניינות לגבי יכולת החיזוי של המשתנים השונים:

- **חיזוי pH:** המודל מפגין יציבות יחסית מרשימה. אף על פי שהשגיאה המוחלטת עומדת על כ-0.48 יחידות, **השגיאה היחסית היא נמוכה מאוד ועומדת על כ-5% בלבד**. נתון זה מעיד כי המודל מצליח ללמוד היטב את הדינמיקה הכימית הבסיסית של המצע, וכי הסטייה האבסולוטית נובעת ככל הנראה מכיול (Calibration) ולא מחוסר יכולת חיזוי.
- **חיזוי EC:** כאן התמונה מורכבת יותר. אמנם השגיאה המוחלטת נראית זניחה כ-0.10 mS/cm אך בשל ערכי המוליכות הנמוכים שנמדדו בתקופה זו, שגיאה זו מתרגמת לסטייה יחסית משמעותית של כ-34%. ממצא זה מדגיש את האתגר בחיזוי ערכים נמוכים (Low-Range Sensitivity) ומסמן את הצורך בשיפור רגישות המודל לשינויים עדינים בריכוזי הדשן בהמשך המחקר.
- **יציבות בזמן:** בביתוח Walk-Forward המודל שמר על עקביות בחיזוי לטווחים קצרים ובינוניים ממצא המחזק את הפוטנציאל לשימוש בו ככלי תכנון עתידי.

ממצאים אלו מדגימים היתכנות לפיתוח מודל חיזוי אפקטיבי כאשר חוזקו הנוכחי הוא בזיהוי מגמות pH יציבות, בעוד האתגר המרכזי להמשך הוא דיוק המודל בערכי EC נמוכים.



2. מבוא:

2.1. רקע כללי: האתגר בבקרת אקלים ובית שורשים

החקלאות המודרנית ניצבת בפני אתגר כפול: הצורך למקסם את היבול ליחידת שטח, תוך צמצום השימוש במשאבים (מים ודשן) ומזעור ההשפעה הסביבתית. בחממות טכנולוגיות, המבוססות לרוב על מצעים מנותקים, השליטה על תנאי בית השורשים (Rootzone) היא המפתח להשגת מטרות אלו. בניגוד לגידולי שדה בקרקע טבעית, למצע המנותק יש באפר קיבול (Buffering Capacity) נמוך, משמעות הדבר היא שכל טעות בדישון או בהשקיה משפיעה באופן מיידי וחריף על הצמח.

בסביבה דינמית זו, שני המדדים הקריטיים ביותר לניטור הם:

- מוליכות חשמלית (EC): מדד המייצג את סך המלחים המומסים בתמיסה. שמירה על רמת EC אופטימלית היא איזון עדין – ערך נמוך מדי יגרום לחוסר הזנה ופגיעה בקצב הצימוח, בעוד שערך גבוה מדי ייצור לחץ אוסמוטי שיקשה על הצמח לקלוט מים, ואף עלול לגרום לצריבות והמתקת יבול מוקדמת.
- חומציות (pH): מדד הקובע את המסיסות והזמינות של יסודות ההזנה (בעיקר מיקרו-אלמנטים כמו ברזל ומנגן). סטייה ב-pH עלולה להוביל למצבי חוסר (Deficiency) או רעילות (Toxicity) גם אם הדשן סופק בכמות הנכונה.

2.2. הפער בין הרצוי למצוי בניטור

על אף החשיבות הקריטית של מדדים אלו, קבלת תמונת מצב אמينة ורציפה של בית השורשים מהווה כיום "צוואר בקבוק" טכנולוגי. המגדל נדרש לבחור בין שתי חלופות שאינן מושלמות:

1. דגימות מעבדה ידניות (Manual Sampling):

זוהי שיטת ה-Gold Standard מבחינת דיוק. דגימת נקז או מיצוי מצע נבדקת במכשיר מעבדה מכויל. עם זאת, שיטה זו דורשת כוח אדם וזמן, ולכן מתבצעת לרוב בתדירות נמוכה (אחת ליום או למספר ימים). התוצאה היא "נקודות עיוורות" בזמן; ייתכן שבין שתי דגימות תקינות התרחש אירוע קיצון של המלחה (Spike) שגרם לנזק בלתי הפיך, אך לא תועד. כמו כן, שיטה זו היא ריאקטיבית (תגובתית) – המידע מתקבל לאחר מעשה.

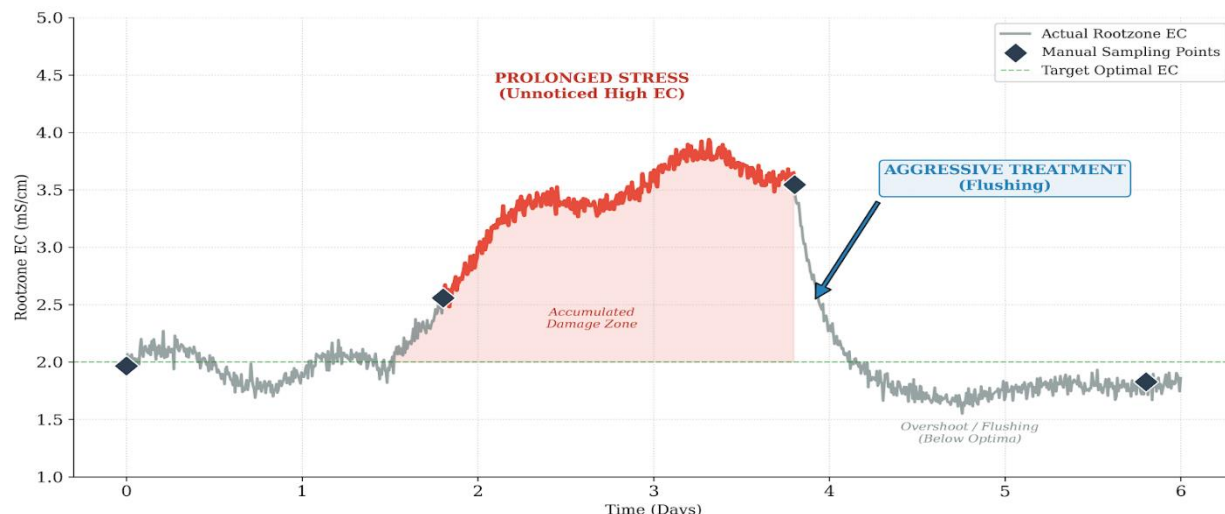


Figure 1: Current Workflow - Manual Sampling with Late Detection and Reactive Correction



2. חיישני קרקע רציפים (In-situ Sensors):

חיישנים אלו מספקים רצף נתונים (Time-Series) ומאפשרים זיהוי מגמות בזמן אמת. אולם, השימוש בהם טומן בחובו אתגרים טכניים משמעותיים:

- סחיפה וכיול (Drift & Calibration): האלקטרודות נוטות לצבור משקעים וביופילם, הגורמים לקריאה "לנדוד" עם הזמן ולספק נתונים שגויים.
- ייצוגיות מרחבית: חיישן מודד נקודה בודדת במרחב. בחממה גדולה, או אפילו בתוך עציץ בודד, קיימת שונות מרחבית גבוהה ברטיבות ובמליחות. חיישן הממוקם ליד הטפטפת יראה ערך שונה לחלוטין מחיישן הממוקם בתחתית המצע.
- עלות: פריסה רחבה של חיישנים איכותיים דורשת השקעה כספית גבוהה ותחזוקה שוטפת.

2.3. הבסיס הפיזיקלי-ביולוגי

הקושי בחיזוי משתני השורשים נובע מכך שבית השורשים הוא מערכת דינמית המושפעת ממאזן מסה (Mass Balance) משתנה. הצמח קולט מים ודשן בקצבים שונים:

- בימים של קרינה גבוהה: הצמח מבצע דיות (Transpiration) מוגברת ו"שותה" בעיקר מים נקיים, מה שמותיר את המלחים במצע וגורם לעליית EC מהירה.
- בימים מעוננים: קצב הדיות יורד, וצריכת המים והדשן מתאזנת. המודל שלנו נועד ללמוד את הדפוסים הללו (הקשר הלא-ליניארי בין קרינה לצריכת מים) ולחזות את הצטברות המלחים שנובעת מהם.

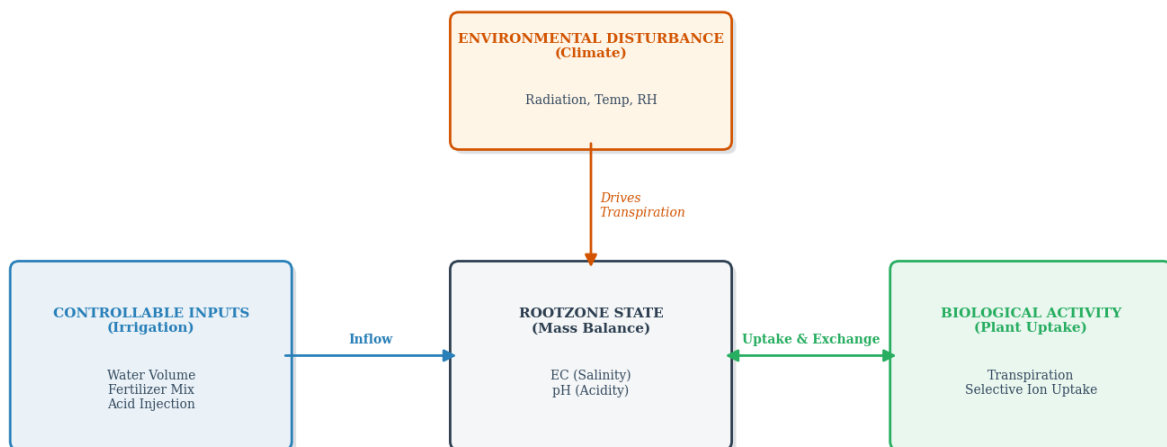


Figure 2: The Physical-Biological Basis - Factors influencing rootzone variables (EC & pH) resulting from the balance between Irrigation, Climate, and Plant Activity.



2.4. למה למידת מכונה?

באופן מסורתי, ניתן לחזות תהליכים אלו באמצעות משוואות פיזיקליות מורכבות. אולם, מודלים אלו דורשים פרמטרים רבים שקשה מאוד למדוד בשטח (כגון מוליכות הידראולית של המצע, נקבוביות משתנה ושטח פנים של השורשים). גישת למידת המכונה (Data-Driven Approach) עוקפת את הצורך במדידת הפרמטרים הללו. במקום "להבין" את הפיזיקה ברמת המיקרו, המודל לומד את ההתנהגות האפקטיבית של המערכת ישירות מתוך הנתונים ההיסטוריים, ובכך מאפשר דיוק גבוה בעלות חישובית נמוכה.

2.5. הפתרון המוצע: "חיישן וירטואלי" מבוסס נתונים

פרויקט זה מציע גישה חדשנית המבוססת על תפיסת ה-Soft Sensing ("חיישן תוכנה"). ההנחה היא שקיים קשר פיזיקלי וכימי הדוק בין תנאי הסביבה החיצוניים (מזג האוויר), פעולות המגדל (השקיה ודישון) והתגובה בבית השורשים.

באמצעות אלגוריתמים של למידת מכונה (Machine Learning) ניתן למדל קשרים מורכבים ובלתי-ליניאריים אלו, ולייצר תחזית מדויקת של מצב בית השורשים על בסיס נתונים זמינים וזולים, ללא תלות בחיישני קרקע יקרים.

2.6. מטרות הפרויקט

1. אינטגרציית נתונים: יצירת מסד נתונים מאוחד (Master Dataset) המנרמל ומשלב מקורות מידע הטרוגניים: נתוני אקלים חיצוני, בקרת השקיה, ודגימות מעבדה.
2. פיתוח מודל חיזוי (Predictive Modeling): בניית מודל למידת מכונה המסוגל לחזות את ערכי ה-EC וה-pH בצעד הזמן הבא בהינתן המצב הנוכחי ותחזית האקלים וההשקיה. המטרה היא לספק למגדל יכולת פרואקטיבית לצפות בעיות לפני שהן מתרחשות.
3. ולידציה בתנאי אמת: בחינת ביצועי המודל בשיטת Walk-Forward Validation המדמה את התקדמות עונת הגידול, כדי להבטיח שהפתרון יציב, אמין וניתן ליישום במערכות תומכות החלטה.



3. מושגים בסיסיים

לצורך הבנת העבודה, להלן הסבר על המושגים המרכזיים בתחומי האגרונומיה ולמידת המכונה בהם נעשה שימוש:

3.1. מושגים אגרונומיים

- מוליכות חשמלית (EC - Electrical Conductivity): מדד לכמות המלחים המומסים בתמיסת הקרקע. בחקלאות, זהו המדד המרכזי לרמת הדישון. ערכים גבוהים מדי עלולים לגרום להמלחה ופגיעה בצמח, וערכים נמוכים מעידים על חוסר בדשן. נמדד ביחידות של mS/cm.
- חומציות (pH): מדד לריכוז יוני המימן בתמיסה. ה-pH קובע את זמינות המיקרו-אלמנטים (כגון ברזל ומנגן) לצמח.
- התאדות-דיות (ET0 - Evapotranspiration): סך המים המתאדים מהקרקע והמופרשים מהצמח (טרנספירציה). זהו מדד קריטי לקביעת מנות ההשקיה, המושפע מקרינה, טמפרטורה, לחות ורוח.
- בית שורשים (Rootzone): נפח הקרקע או המצע שבו מתפתחת מערכת השורשים של הצמח. התנאים באזור זה משפיעים ישירות על קצב גדילת הצמח ובריאותו.

3.2. מושגים בלמידת מכונה

- מודל רגרסיה (Regression Model): מודל שנועד לחזות ערך מספרי רציף (בניגוד לסיווג לקטגוריות). בפרויקט זה אנו חוזים ערכים רציפים של משתני מיקרו אקלים, EC ו-pH.
- XGBoost: אלגוריתם מתקדם מבוסס עצי החלטה (Gradient Boosting), הידוע ביעילותו ובמהירותו. שימש בפרויקט להשלמת נתוני אקלים חסרים ובחיזוי משתני בית השורשים.
- LightGBM (Light Gradient Boosting Machine): אלגוריתם למידה מתקדם המבוסס על שיטת Gradient Boosting (שיפור הדרגתי של עצי החלטה). האלגוריתם מתאפיין בשיטת גדילת עץ ייחודית (Leaf-wise) המאפשרת מהירות אימון גבוהה במיוחד, ניצול יעיל של זיכרון ודיוק רב, והוא נחשב לאידיאלי לעיבוד סדרות עיתיות ונתונים בהיקף רחב. שימש בפרויקט לחיזוי ערכי מיקרו אקלים עתידיים בתוך החממה על בסיס האקלים החיצוני.
- Walk-Forward Validation: שיטת אימות למודלים של סדרות עיתיות (Time Series) במקום חלוקה אקראית ל"אימון" ו"מבחן", אנו מאמנים את המודל על נתוני העבר עד נקודה מסוימת, בודקים על העתיד הקרוב, ואז "מזיזים את החלון" קדימה. שיטה זו מדמה במדויק כיצד המערכת תפעל במציאות, יום אחר יום.
- הנדסת פיצ'רים (Feature Engineering): תהליך יצירת משתנים חדשים מתוך הנתונים הגולמיים כדי לשפר את יכולת הלמידה של המודל. דוגמאות מפרויקט זה: "זמן שעבר מאז ההשקיה האחרונה", "סכום הקרינה המצטבר ב-24 שעות אחרונות".



4. סקירת ספרות

השימוש במודלים מבוססי נתונים (Data-Driven Models) בחקלאות מדייקת צבר תאוצה משמעותית בשנים האחרונות. מחקרים שונים הראו כיצד ניתן להחליף או לתגבר מדידות פיזיקליות יקרות באמצעות אלגוריתמים המעבדים נתוני סביבה ותפעול.

אחד האתגרים המרכזיים בגידול בחממות הוא השליטה ברמת המליחות (EC) בבית השורשים. במחקרם של **Moon et al. (2018)** בחנו החוקרים את היכולת לחזות את ה-EC של תמיסת ההזנה במערכות הידרופוניות סגורות (Closed-loop soilless cultures). החוקרים השתמשו ברשת נוירונים מסוג RNN (Recurrent Neural Network) ובפרט במודל LSTM, הידוע ביכולתו ללמוד סדרות עיתיות. המחקר הראה כי שילוב של **משתני אקלים** (קרינה, טמפרטורה ולחות יחסית) יחד עם **נתוני תפעול** (כמות השקיה וריכוז דשן במי ההשקיה), מאפשר לחזות בדיוק רב את ה-EC העתידי במצע. המסקנה המרכזית ממחקר זה, הרלוונטית לעבודתנו, היא שקיים קשר הדוק וניתן למידול בין עומס האקלים החיצוני לבין דינמיקה הכימית בבית השורשים, וכי היסטוריית הנתונים (Time-lagged features) היא קריטית לביצועי המודל.

בעוד ש-Moon- התמקד בכימיה של המים **Lü et al. (2025)**, הציגו גישה מתקדמת לחיזוי רטיבות קרקע בבית השורשים (RZSM – Root Zone Soil Moisture). המחקר התמודד עם הקושי למדוד רטיבות בעומק הקרקע והציע מודל היברידי המשלב מודל פיזיקלי הידרולוגי (Hydrus-1D) עם מודל למידה עמוקה מסוג CNN-LSTM-Attention. המחקר הדגים כיצד מודלים של למידה עמוקה (Deep Learning) יכולים ללמוד את הקשרים הלא-ליניאריים המורכבים בין רטיבות פני השטח (הניתנת למדידה קלה או חישה מרחוק) לבין המתרחש בעומק בית השורשים. השימוש במנגנון קשב (Attention Mechanism) אפשר למודל להתמקד בטווחי הזמן המשפיעים ביותר על המצב הנוכחי. אף על פי שמחקר זה בוצע בתנאי שדה (Oasis field) ולא בחממה, הוא מחזק את ההנחה כי ניתן לייצר "חיישן וירטואלי" לתנאי עומק הקרקע על בסיס נתונים סביבתיים, וכי שילוב של ארכיטקטורות המביאות בחשבון את המימד הרציף של הזמן כגון LSTM או נגזרותיו משפר משמעותית את הדיוק בהשוואה למודלים סטטיים.

הספרות הקיימת מצביעה על היתכנות גבוהה לשימוש בלמידת מכונה לניטור בית השורשים. בעוד ש-Moon et al. התמקדו ב-EC, Lü et al. ב-רטיבות, פרויקט זה שואף לשלב גישות אלו ולפתח מודל אחוד החוזה **הן את ה-EC והן את ה-pH** תוך שימוש באלגוריתמים של למידת מכונה מבוססי עצים המציעים איזון בין דיוק ליכולת פרשנות ומתאימים ליישום במערכות בקרה תפעוליות בחממות מסחריות.

ביבליוגרפיה

1. **Moon, T., Ahn, T. I., & Son, J. E. (2018).** Forecasting Root-Zone Electrical Conductivity of Nutrient Solutions in Closed-Loop Soilless Cultures via a Recurrent Neural Network Using Environmental and Cultivation Information. *Frontiers in Plant Science*, 9, 859.
2. **Lü, X., Nurmamet, I., Xiao, S., Zhao, J., Yu, X., Aili, Y., & Li, S. (2025).** Spatial-temporal simulation and prediction of root zone soil moisture based on Hydrus-1D and CNN-LSTM-attention models in Yutian Oasis, southern Xinjiang, China. *Pedosphere*, 35(5), 846-857.



5. הגדרת הבעיה

5.1. ניסוח הבעיה

בגידולים אינטנסיביים על מצע מנותק, בית השורשים (Rootzone) מהווה "קופסה שחורה" דינמית. הבעיה המרכזית איתה מתמודד המחקר נובעת מהפער הקיים בין הצורך בבקרת רציפה ומדויקת לבין אמצעי הניטור הקיימים כיום:

1. מגבלת הניטור המקוטע: בדיקות מעבדה (ניקוז/מצע) הן מדויקות אך מספקות תמונת מצב נקודתית ("Snapshot") ודלילה בזמן. בין דגימה לדגימה, המצע עשוי לעבור תנודות חריפות ברמת המליחות (EC) והחומציות (pH) כתוצאה משינויי אקלים או תקלות דישון, ללא ידיעת המגדל.
2. אמינות חיישנים רציפים: חיישנים המוצבים בתוך המצע (In-situ) נוטים לסבול מבעיות כיוול, סחיפה (Drift) ורעש אלקטרוני. בנוסף, הם מודדים נקודה בודדת במרחב ומתקשים לייצג את המצב הממוצע במצע כולו, הסובל משונות מרחבית גבוהה.
3. ניהול תגובתי (Reactive Management): בהיעדר יכולת חיזוי, המגדל נאלץ לפעול בגישה "כיבוי שריפות" תיקון ערכי ה-EC או ה-pH רק לאחר שכבר חרגו מהטווח הרצוי והנזק הפוטנציאלי לצמח כבר החל.

5.2. שאלת המחקר:

האם ניתן לפתח מודל למידת מכונה (Machine Learning) המסוגל לחזות בדיוק אגרונומי את ערכי המוליכות החשמלית (EC) והחומציות (pH) בבית השורשים לטווח הזמן הקרוב, בהתבסס על נתוני אקלים, משטר השקיה ודישון ומדידות עבר?

מתוך שאלה זו נגזרות מספר שאלות משנה:

1. **דיוק:** מהי רמת הדיוק המקסימלית שניתן להשיג בחיזוי (נמדד ב MAE/RMSE) והאם היא מספקת לקבלת עזר החלטה תפעולי?
2. **חשיבות המשתנים (Feature Importance):** אילו משתנים משפיעים בצורה החזקה ביותר על השינויים בבית השורשים? (למשל: האם קרינה משפיעה יותר מכמות המים? האם ההיסטוריה של הדישון חשובה יותר מהטמפרטורה הרגעית?).
3. **יציבות:** האם המודל מסוגל לשמור על יציבות בחיזוי לאורך זמן (Walk-Forward) בתנאי אמת משתנים של עונת הגידול?

5.3. השערת המחקר:

ההשערה היא שקיים קשר דטרמיניסטי (גם אם לא ליניארי ומורכב) בין עומס האקלים החיצוני הקובע את קצב הדיות (Transpiration), אספקת המים והדשן, ושינוי הריכוזים במצע.

אנו משערים כי אלגוריתמים מבוססי עצים כגון Random Forest או XGBoost יצליחו ללמוד את דפוסי "התגובה וההשהיה" של המערכת, ולגשר על הפער שבין המדידות הידניות, ובכך לשמש כ"חיישן וירטואלי" אמין.



6. גישת הפתרון

כדי להתמודד עם האתגר של חיזוי תנאים ביו-כימיים מורכבים בבית השורשים, פיתחנו גישה מערכתית המבוססת על ארכיטקטורה דו-שלבית (Two-Stage Pipeline). גישה זו מפרקת את הבעיה לשני תתי-מודלים, כאשר הפלט של הראשון משמש כקלט לשני. מבנה זה מאפשר לנו להשתמש בנתוני אקלים חיצוניים (הזמינים תמיד) כדי לחזות את המתרחש בתוך המצע.

6.1 מסד הנתונים

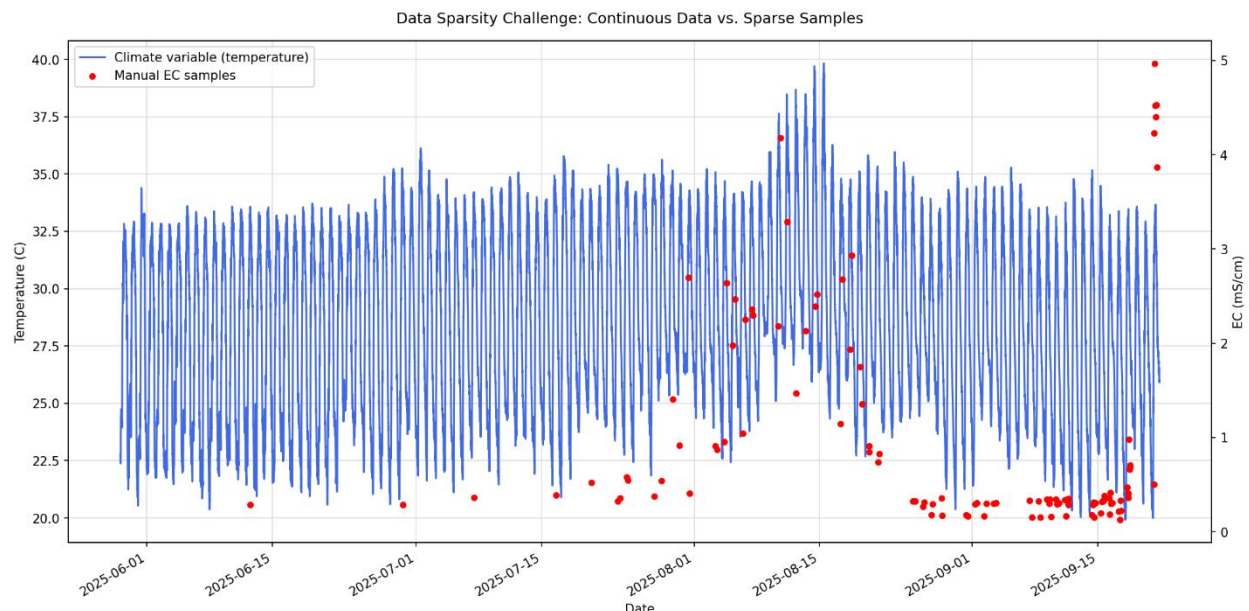
המחקר מתבסס על נתונים אגרו-מטאורולוגיים שנאספו בחממת המחקר במהלך עונת הגידול 2025 (בחודשים מאי-ספטמבר). לצורך אימון ובדיקת המודלים, נבנה מסד נתונים מאוחד המשלב מקורות מידע הטרוגניים, ועבר תהליכי ניקוי וסנכרון. הנתונים נחלקים לשתי קטגוריות עיקריות, הנבדלות ברזולוציית הזמן שלהן:

1. נתונים רציפים (High-Frequency Data):

- מקור: תחנה מטאורולוגית (בית דגן) ולוגר נתונים פנימי בחממה.
- היקף: כ-16,000 רשומות זמן.
- רזולוציה: דגימה כל 10 דקות באופן רציף.
- משתנים: קרינה גלובלית, טמפרטורה, לחות יחסית, מהירות רוח וחישוב התאדות פוטנציאלית (ET0).

2. נתונים בדידים (Low-Frequency Data):

- מקור: דגימות מעבדה ידניות של מי נקז ומצע.
- היקף: כ-100 דגימות תקינות של pH ו-EC.
- האתגר: נתונים אלו מאופיינים בדלילות (Sparsity) גבוהה ובמרווחי זמן לא קבועים (בין יום ל-5 ימים בין דגימות). עובדה זו היוותה את האתגר האלגוריתמי המרכזי בפרויקט, שכן היחס בין משתני הקלט (אקלים) למשתני המטרה (מעבדה) עומד על כ-1:200.





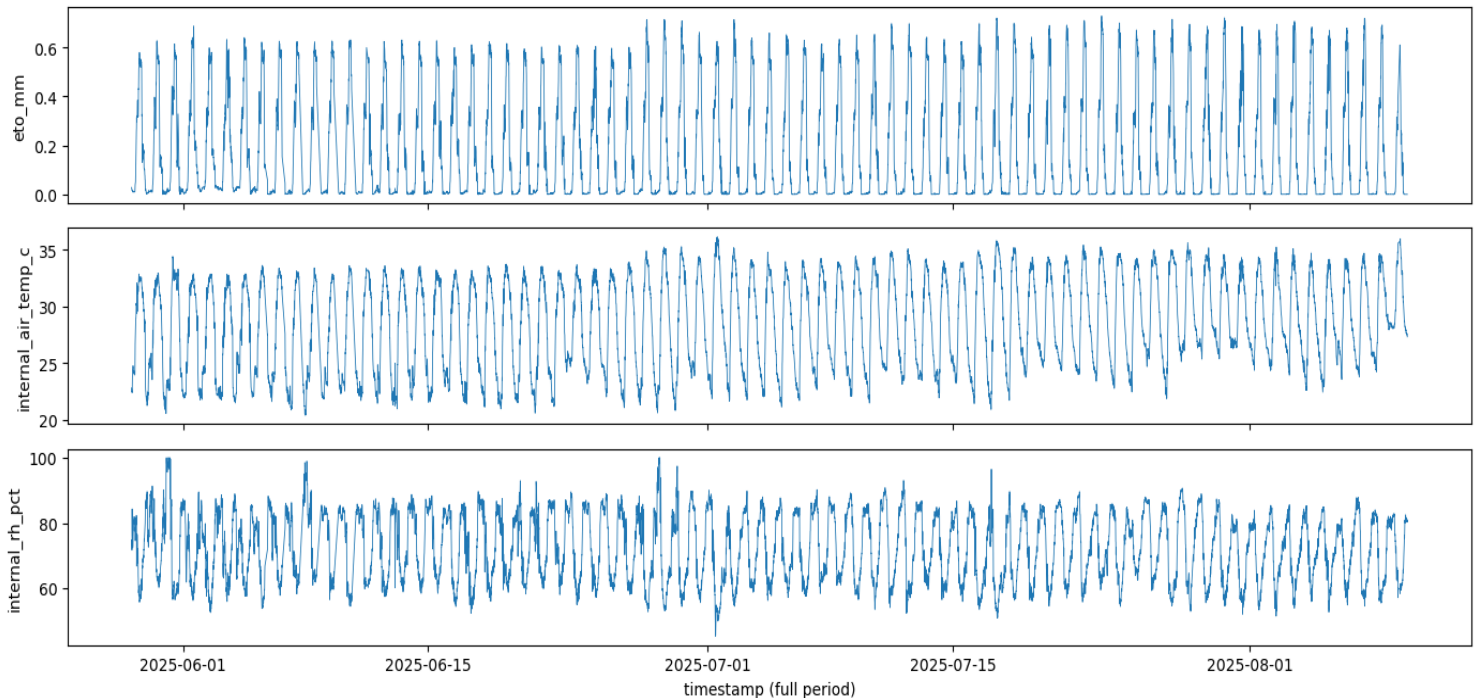
6.2. השלמת נתונים היסטוריים (Data Imputation using XGBoost):

אחד האתגרים המשמעותיים בפרויקט היה היעדר רציפות בנתוני האקלים הפנימי. חיישן הנתונים (Data Logger) הותקן בחממה רק במהלך העונה, כך שעבור התקופה הראשונה של הניסוי היו בידינו נתוני מעבדה (EC/pH) ונתוני חוץ (בית דגן), אך ללא נתוני אקלים פנימי תואמים.

כדי לא לאבד את דגימות המעבדה היקרות מתקופה זו, פיתחנו מודל רגרסיה ייעודי מבוסס XGBoost להשלמת הנתונים החסרים. המודל אומן על התקופה המאוחרת (בה היו נתונים מלאים) כדי ללמוד את הקורלציה בין האקלים החיצוני לפנימי ($R^2 > 0.95$), ולאחר מכן הופעל "לאחור" כדי לשחזר בדיוק רב את הטמפרטורה והלחות הפנימית עבור התקופה שקדמה להתקנת החיישן. פעולה זו אפשרה לנו להגדיל את סט האימון האפקטיבי בכ-30%.

Microclimate Model Performance Results

Target Variable	MAE	R ² Score (%)
EtO (mm)	0.0211	94.56%
Internal Air Temp (°C)	0.3832	98.02%
Internal RH (%)	2.3773	90.08%





6.3. עיבוד מקדים :

בוצע תהליך ניקוי רעשים (Noise Reduction) לסינון ערכים לא פיזיקליים מהחיישנים. כמו כן, בוצע סנכרון זמנים (Time Synchronization) כדי להצמיד לכל דגימת מעבדה בודדת את היסטוריית האקלים וההשקיה המדויקת שהובילה אליה ("חלון השפעה").

6.4. סביבת הפיתוח

מערכת החיזוי פותחה בסביבת Python 3.13 תוך שימוש בספריות המרכזיות הבאות:

- **Pandas** לטיפול ועיבוד סדרות עתיות.
- **Scikit-Learn** למימוש אלגוריתמי ML קלאסיים (Random Forest), חלוקת דאטה-Train-Test Split ומטריקות הערכה.
- **Gradient Boosting** יעילה ששימשה למידול המיקרו-אקלים ובית השורשים.
- **Matplotlib/Seaborn** לוויזואליזציה וניתוח נתונים. הפיתוח בוצע על גבי מחברות Jupyter Notebooks המאפשרות מחקר אינטראקטיבי ותיעוד התהליך.

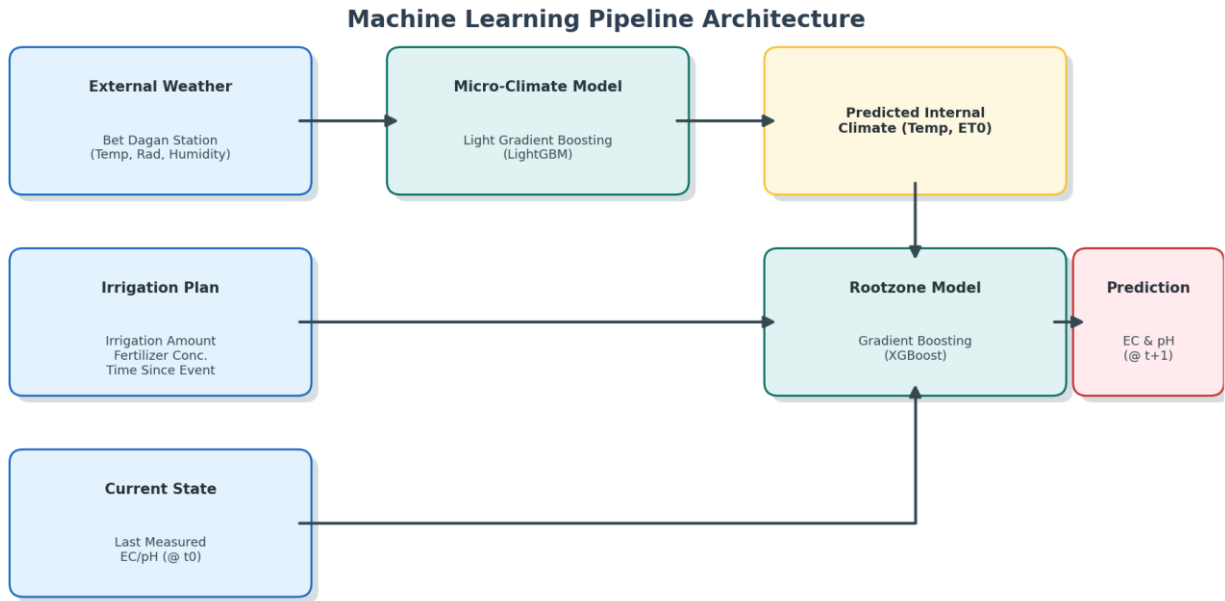
6.5. ארכיטקטורת המערכת

המערכת בנויה משני מודלים עוקבים המזינים זה את זה:

- שלב 1: מודל "תרגום" אקלים (Micro-Climates Model):
מודל זה משמש כ"מתווך". הוא מקבל כקלט את תחזית מזג האוויר החיצונית (תחנת בית דגן) וחוזר את תנאי המיקרו-אקלים בתוך החממה (טמפרטורה פנימית, לחות יחסית, קרינה ET0).
- **הרציונל:** הצמח ובית השורשים מושפעים מהתנאים בתוך המבנה, השונים משמעותית מהתנאים בחוץ בשל אפקט החממה. שלב זה מאפשר לנו להשתמש בתחזיות מטאורולוגיות סטנדרטיות כבסיס לחיזוי.
- **אלגוריתם LightGBM (Light Gradient Boosting Machine):** הבחירה ב-LightGBM נעשתה בשל יתרונותיו המובהקים בעיבוד סדרות עתיות ובמהירות האימון הגבוהה שלו בהשוואה לאלגוריתמים מסורתיים. המודל אומן ללמוד את פונקציית המעבר הלא-ליניארית שבין הקרינה, הלחות והטמפרטורה בחוץ, לבין הטמפרטורה, הלחות וההתאדות (ET0) בתוך המבנה המוגן.
- שלב 2: מודל בית השורשים (Rootzone Dynamics Model):
זהו מודל הליבה. הוא מקבל את המיקרו-אקלים החזוי (משלב 1), את תוכנית ההשקיה והדישון, ואת המצב הנוכחי וחוזר את ערכי ה-EC וה-pH בצעד הזמן הבא.



- **אלגוריתם XGBoost Regressor** : אלגוריתם זה נבחר בשל עמידותו לרעש (Noise Robustness) ויכולתו להתמודד היטב עם מערכי נתונים קטנים יחסית (כ-100 דגימות ידניות) מבלי להיכנס לבעיית התאמת-יתר (Overfitting).



6.6. אסטרטגיית איחוד הנתונים

אחד האתגרים המרכזיים בפרויקט היה השוני ברזולוציית הזמן של הנתונים:

- **נתונים "צפופים"**: נתוני אקלים ולוגר נתונים (כל 10 דקות).
- **נתונים "דלילים"**: דגימות מעבדה ידניות (אחת ליום או למספר ימים רוב הזמן).

הפתרון: יצירת Master Dataset ברזולוציה אחידה של 10 דקות.

הנתונים הדלילים "הוצמדו" לנקודות הזמן הקרובות ביותר (Nearest Neighbor Interpolation), אך האימון של מודל בית השורשים בוצע רק על גבי "חלונות זמן" שיש בהם דגימת אמת בעוד החיזוי יכול להתבצע באופן רציף.



6.7. הנדסת פיצ'רים (Feature Engineering)

כדי לאפשר למודל "להבין" את התהליכים הביולוגיים והפיזיקליים, הנתונים הגולמיים עובדו למשתנים חדשים בעלי משמעות (Domain Knowledge):

- **משתני צבירה (Cumulative Features):** במקום להסתכל רק על הקרינה הרגעית, המודל מחשב את "סך הקרינה ב-24 השעות האחרונות" או "סך ההשקיה ב-8 השעות האחרונות". משתנים אלו מייצגים את ה"זיכרון" של המערכת.
- **זמן מאז אירוע (Time-Since Events):** משתנים המודדים כמה זמן עבר מאז ההשקיה או הדישון האחרון. משתנים אלו קריטיים לזיהוי תבניות של עלייה וריכוז מלחים.
- **מצב הצמח (Canopy Cover):** שילוב נתון כיסוי הנוף כאינדיקציה לגודל הצמח וליכולת שלו לקלוט מים ודשן (Transpiration rate).

6.8. אסטרטגיית אימון ובדיקה (Walk-Forward Validation)

מכיוון שמדובר בסדרה עיתית (Time Series), לא ניתן להשתמש בחלוקה אקראית רגילה (Random Split) לסט אימון וסט מבחן, מחשש לזליגת מידע (Data Leakage) מהעתיד לעבר.

במקום זאת, יושמה שיטת Walk-Forward Validation המדמה את התקדמות הזמן.

בפרויקט זה הגדרנו שני פרוטוקולים שונים של אופק חיזוי (Forecast Horizon) בהתאם לסוג המודל וזמינות הנתונים:

1. עבור מודל המיקרו-אקלים (נתונים רציפים):

המודל אומן על נתונים היסטוריים וביצע חיזוי של 24 שעות קדימה בכל איטרציה.

○ **הרציונל:** לאפשר למגדל לתכנן את משטר ההשקיה ליום המחרת במלואו על בסיס התחזית.

2. עבור מודל בית השורשים (נתונים בדידים/דלילים):

המודל אומן על כל הדגימות הידניות עד נקודה t , וביצע חיזוי של דגימה אחת קדימה $t + 1$.

○ **הרציונל:** מכיוון שמדידות האמת (Ground Truth) של EC ו-pH מגיעות מהמעבדה באופן לא סדיר, המודל נבחן ביכולתו לחזות את תוצאת המעבדה הבאה בתור, ללא תלות במשך הזמן שעבר (שעות או ימים).

לאחר כל חיזוי, חלון האימון הורחב לכלול את הנתונים החדשים ("Real-Time Update") והתהליך חזר חלילה עד סוף תקופת הניסוי.



7. תוצאות ראשוניות ואתגרים

7.1. תוצאות המודלים

נכון לשלב זה של המחקר, המערכת נבחנה על נתונים שנאספו במשך כמספר חודשים (מיוני עד ספטמבר). להלן הביצועים של שני המודלים המרכזיים:

א. מודל המיקרו-אקלים:

מודל זה הציג ביצועים יוצאי דופן והוכיח כי ניתן לחזות את האקלים הפנימי בדיוק רב על בסיס נתונים חיצוניים בלבד.

- טמפרטורה פנימית: התקבל מקדם מתאם R^2 של 0.97 עם שגיאה ממוצעת של פחות מ- 0.8° צלזיוס.
- לחות יחסית: הדיוק עמד על R^2 של 0.87 עם שגיאה ממוצעת של 2.5% בלבד וטעות יחסית של 3.5%, נתון קריטי לחיזוי התאדות.
- התאדות (ET0): התקבל דיוק גבוה במיוחד עם מקדם מתאם R^2 של 0.97, ושגיאה מוחלטת ממוצעת (MAE) של 0.016 מ"מ בלבד לכל מרווח זמן של 10 דקות. דיוק זה הוא קריטי שכן ה-ET0 מהווה את הבסיס לחישוב מנות ההשקיה.





Model Performance Results

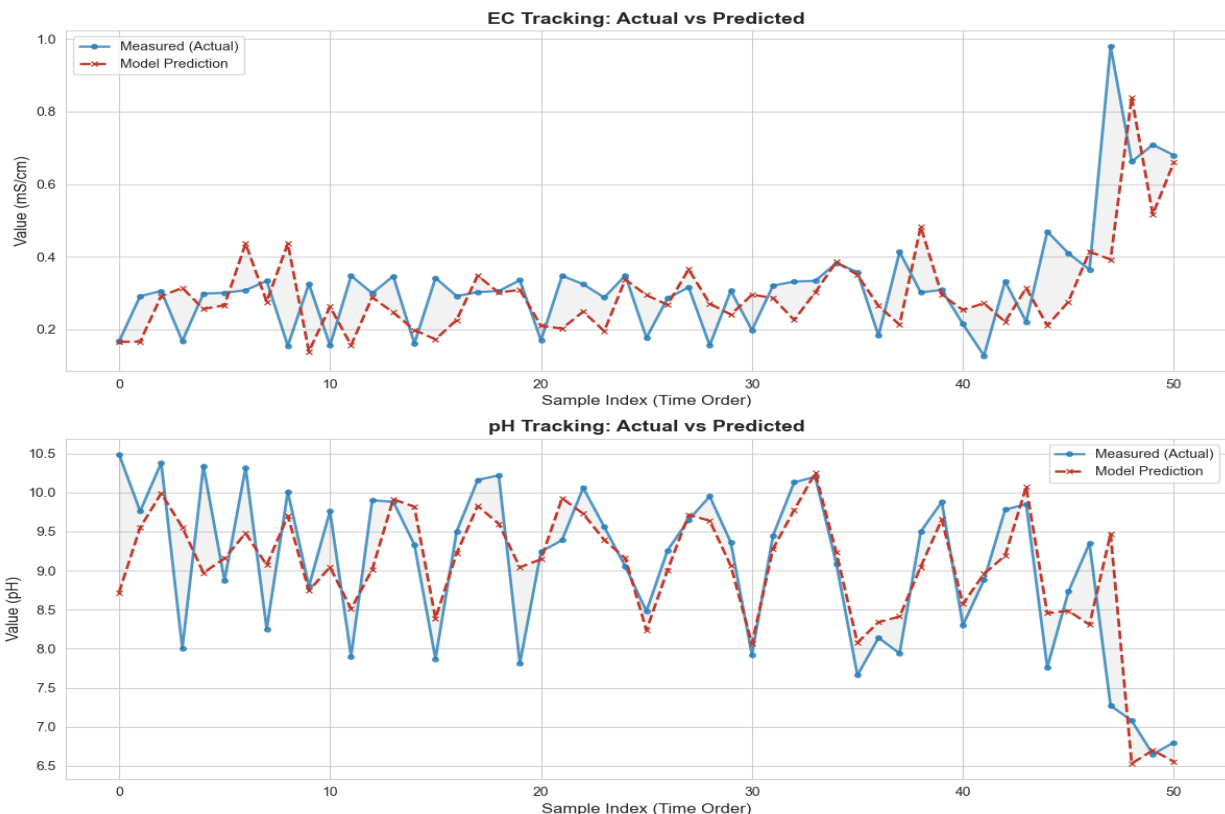
Target Variable	MAE	RMSE	R ² Score
Internal Air Temp (°C)	0.3863	0.5062	97.86%
Internal Radiation (W/m ²)	20.5592	39.1329	96.89%
Evapotranspiration (mm)	0.0158	0.0296	97.18%
Internal Relative Humidity (%)	2.5209	3.2580	86.77%

- משמעות: המודל משמש כ"תחנה מטאורולוגית וירטואלית" אמינה מאוד עבור המודל הבא בשרשרת, ויכול לחזות בצורה אמינה את האקלים הפנימי בחממה לטווח של 24 שעות קדימה.

ב. מודל בית השורשים:

כאן התוצאות משקפות את המורכבות הביולוגית של המערכת:

- חיזוי pH: המודל הציג יציבות מרשימה בזיהוי המגמה, עם שגיאה יחסית נמוכה של כ-5% (MAE 0.48) זהו הישג משמעותי בהינתן שה pH מושפע מתהליכים כימיים זעירים בקרקע.
- חיזוי EC: המודל הצליח לחזות את ערכי המוליכות בשגיאה אבסולוטית נמוכה של 0.10 mS/cm. עם זאת, בשל ערכי המליחות הנמוכים שנמדדו בתקופה זו, השגיאה היחסית נותרה גבוהה (כ-34%), אתגר המפורט בהמשך.





Overall Next-Observation Performance (with Relative Error)

Target Variable	MAE	RMSE	Total Rel. Error
pH	0.4820	0.6635	5.46%
EC (mS/cm)	0.0996	0.1393	33.95%

7.2. אתגרים ודרכי פתרון

במהלך הפיתוח נתקלנו בשלושה אתגרים מרכזיים הנובעים מאופי הנתונים והמערכת. להלן הניתוח והפתרונות:

7.2.1. אתגר 1: "הקופסה השחורה" – העדר ניטור רציף

- **הבעיה:** בחממה הנחקרת לא קיים כלל חיישן לניטור רציף של EC ו-pH המשמעות היא שהמערכת נמצאת ב"עיוורון" מוחלט רוב הזמן. ישנן נקודות אמת בודדות (דגימות מעבדה) המרוחקות זו מזו בזמן (לעיתים ימים שלמים), ואין שום דרך לדעת מה קרה לבית השורשים בתווך שביניהן – האם הייתה קפיצת מליחות? האם ה-pH צנח לאחר הדישון?
- **הפתרון:** פיתוח "חיישן וירטואלי" טהור, המודל תוכנן כך שהוא אינו מסתמך על משוב (Feedback) רציף מהקרקע, אלא בונה מחדש את "סיפור" בית השורשים אך ורק על סמך **הקלט** (כמויות המים והדשן שנכנסו) ו**הסביבה** (מיקרו אקלים). זוהי משימה מורכבת יותר מכיול חיישן קיים, שכן המודל נדרש לסמלך את התהליך הכימי כולו.

7.2.2. אתגר 2: בעיית הדלילות

- **הבעיה:** מאחר שאין חיישנים, ה-Ground Truth שלנו מוגבל לכ-100 דגימות ידניות בלבד שנאספו לאורך מספר חודשים. אימון מודל למידת מכונה הדורש בדרך כלל (Big Data) על סט נתונים כה דליל יוצר סיכון גבוה ל-Overfitting (שינון הנתונים).
- **הפתרון:**

1. **הנדסת פיצ'רים אגרסיבית:** במקום להזין למודל נתונים רגועים, יצרנו משתנים "היסטוריים" (כגון: "סך הדשן שהצטבר ב-48 השעות האחרונות" או "זמן שחלף מאז ההשקיה האחרונה"). משתנים אלו מפצים על חוסר הנתונים בכך שהם נותנים למודל הקשר (Context) רחב לכל דגימה.

2. **שימוש באלגוריתם XGBoost:** נבחר אלגוריתם המבוסס על אננסמבל של עצים ומוגבל בעומקו, אשר הוכיח עמידות גבוהה בעבודה עם דגימות מעטות.

7.2.3. אתגר 3: רגישות המודל בערכי קיצון נמוכים

- **הבעיה:** במהלך תקופת הניסוי נמדדו בבית השורשים הרבה ערכי EC נמוכים במיוחד, בטווח של 0.2-0.5 mS/cm (ערכים האופייניים למצע שטוף כמעט לחלוטין). בטווח דינמי זה, אנו נתקלים בבעיה מתמטית ופיזיקלית כפולה:



1. הטיה סטטיסטית : מבחינה מתמטית, שגיאת חיזוי אבסולוטית של 0.1 mS/cm נחשבת מצוינת. כאשר ה-EC הוא 2.5, מדובר בטעות של 4% בלבד. אולם כאשר ה-EC הוא 0.2, אותה סטייה בדיוק מתורגמת לטעות יחסית של 50%. הדבר יוצר עיוות במדדי הערכת המודל.

2. יחס אות לרעש: בערכים כה נמוכים, ה"אות" (השינוי האמיתי בריכוז המלחים כתוצאה מדישון) הוא חלש מאוד, וקשה להפרדה מ"רעש הרקע" (טעויות מדידה קטנות, השפעת טמפרטורה על המוליכות, או שאריות מלחים במים). המודל מתקשה לזהות דפוסים ברורים כאשר השינויים הם בסדרי גודל מיקרוסקופיים.

המשמעות האגרנומית: חשוב לציין כי מבחינה מעשית ההבדל בין EC של 0.2 ל-0.3 הוא זניח עבור הצמח – שניהם מצביעים על חוסר דשן ודורשים אותה פעולה מתקנת. לכן, הטעות הגבוהה באחוזים אינה מעידה בהכרח על כישלון המודל ככלי תומך החלטה.

• הפתרון: טרנספורמציה לוגריתמית של המטרה (Log-Target Transformation)

משתנים כימיים כמו ריכוזים (EC) נוטים להתפלג בצורה שאינה נורמלית, והשונות (Variance) של השגיאה גדלה ככל שהערך גדל.

במקום לאמן את המודל לחזות את הערך הגולמי (y) נאמן אותו לחזות את הלוגריתם של הערך ($\log(y)$). טרנספורמציה זו "מותחת" את הטווחים הנמוכים ומכווצת את הגבוהים. כאשר ממירים את התחזית חזרה על ידי שימוש ב-Exponent המודל נוטה באופן טבעי להימנע מחיזוי ערכים שליליים או אפסיים לא הגיוניים, ומשפר את הדיוק היחסי בערכים הנמוכים.

7.2.4. אתגר 4: אי-סדירות זמנית של הדגימות

• **הבעיה:** בניגוד לנתוני אקלים המגיעים בקצב קבוע (כל 10 דקות), דגימות המעבדה נלקחות באופן לא סדיר - לעיתים ההפרש הוא כמה שעות, לפעמים יום אחד, ולעיתים 4-5 ימים (סופי שבוע/חגים). מצב זה מקשה על שימוש באלגוריתמים קלאסיים של סדרות עתיות כגון ARIMA או LSTM המניחים צעדי זמן קבועים ($t, t+1, t+2$). המודל נדרש ללמוד לחזות את המצב הבא, בלי לדעת מראש כמה זמן יעבור עד אליו.

• **הפתרון:** שימוש בפיצ'ר של "זמן יחסי" הזנו למודל משתנה מפורש המייצג את מספר השעות שעברו מאז הדגימה האחרונה, כך שהוא למד לשקלל את "דעיכת המידע" – ככל שעבר יותר זמן, עליו להסתמך פחות על המצב ההתחלתי ויותר על האינטגרציה של ההשקיה והאקלים בתווך.

7.2.5. אתגר 5: אפקט הזיכרון ותלות במצב ההתחלתי

• **הבעיה:** תגובת המצע לדישון אינה ליניארית ואינה אחידה, אלא תלויה בהיסטוריה ובמצב הנוכחי שלו.

1. **אפקט הזיכרון:** השקיה בודדת במים מתוקים לא מורידה את ה-EC-מיד, אלא מבצעת שטיפה הדרגתית.



2. **תלות במצב (State Dependency):** אותה מנת דשן בדיוק תוביל לתוצאה כימית שונה לחלוטין אם היא ניתנת למצע שכבר רווי במלחים וחומצי, לעומת מצע שטוף ובסיסי. התעלמות מהמצב ההתחלתי גורמת לשגיאות חיצוני גסות.

3. **אפקט המהילה:** השפעת הדשן נגזרת מהריכוז הסופי בתמיסה, ולא רק מהכמות המוחלטת (גרמים). השקיה בנפח מים גדול "מדללת" את הדשן וממתנת את השפעתו על ה-EC.

• **הפתרון:** הנדסת פיצ'רים מרובדת: (Multi-Layer Feature Engineering)

1. **חלונות צבירה (Accumulation Windows):** יצירת משתנים מסכמים לחלונות זמן שונים (8, 16, 24 שעות) כדי לאפשר למודל לזהות דפוסים של הצטברות מלחים או מחזורי שטיפה.

2. **יחסי ריכוז (Concentration Ratios):** במקום להשתמש רק בכמויות מוחלטות של דשן (mg), יצרנו פיצ'רים המחשבים את היחס בין כמות הדשן לנפח המים באותה השקיה. משתנים אלו מייצגים את הריכוז בפועל (ppm) ואת אפקט המהילה, ומסייעים למודל לחקות את הפיזיקה האמיתית של תמיסת הקרקע.

7.2.6. אתגר 6: אי-רציפות בנתוני האקלים

• **הבעיה:** מסד הנתונים של דגימות המעבדה (EC/pH) החל להיבנות עוד לפני שהותקנה בחממה מערכת הניטור הרציפה (Data Logger). נוצר מצב שבו עבור חלק נכבד מדגימות הקרקע המוקדמות, לא היה קיים תיעוד מקביל של האקלים הפנימי (טמפרטורה/לחות), מה שמנע את האפשרות להשתמש בהן לאימון המודל.

• **הפתרון:** שחזור נתונים לאחור. במקום לוותר על הדגימות היקרות, פיתחנו מודל רגרסיה ייעודי מבוסס XGBoost שאומן על התקופה המאוחרת יותר, בה היו נתונים מלאים. המודל למד את המתאם הגבוה שבין נתוני התחנה המטאורולוגית החיצונית (בית דגן) לבין האקלים בתוך החממה. לאחר שהמודל הגיע לדיוק גבוה ($R^2 > 0.95$), השתמשנו בו כדי לשחזר ("לחזות לאחור") את תנאי האקלים הפנימיים עבור התקופה החסרה. פעולה זו אפשרה לנו להציל עשרות דגימות מעבדה ולהגדיל את סט האימון האפקטיבי בכ-30%.



8. תוכנית עבודה ולוח זמנים

תוכנית העבודה להמשך המחקר מתמקדת במאמץ העיקרי: **מקסום הדיוק של מודל ה-EC וה-pH**. תהליך זה יתבצע בצורה איטרטיבית (ניסוי וטעיה מובנה) של הוספת משתנים, סינון רעשים וכיול האלגוריתם. במקביל, יבוצעו ניתוחים משלימים ("מעטפת") כדי להדגים את יציבות המערכת ולייצר תוצרים ויזואליים לדוח הסופי.

8.1. שלב 1: הנדסת פיצורים ושיפור המודל (המאמץ העיקרי)

המטרה: הורדת שגיאת החיזוי (MAE) באופן עקבי על ידי טיוב הנתונים הנכנסים. זהו לב העבודה בחודשים הקרובים.

• משימה 1.1 - בחינת משתני מגמה (Trend Features)

- יצירת משתנים חדשים המייצגים את קצב השינוי (נגזרת) של ה-pH וה-EC בדגימות האחרונות.
- בדיקה האם הוספת מידע על "כיוון המגמה" (עלייה/ירידה) משפרת את החיזוי יותר מאשר הערך האבסולוטי בלבד.

• משימה 1.2 - ניסוי חלונות זמן (Time Windows Optimization)

- ביצוע סדרת ניסויים כדי למצוא את "חלון הזיכרון" האופטימלי של המצע: האם המודל מדויק יותר כשהוא מסתכל על צבירת דשן של 12 שעות? 24 שעות? 48 שעות?
- אופטימיזציה של משתני השהיה – (Lag Features) כמה זמן לוקח להשקיה להשפיע על המדידה?

• משימה 1.3 - יחסי גומלין (Interaction Features)

- בדיקת פיצורים מורכבים המשלבים מספר גורמים, למשל: יחס דשן-למים (ריכוז בפועל), או מכפלת קרינה בכיסוי נוף (צריכה פוטנציאלית).

8.2. שלב 2: אופטימיזציה של האלגוריתם

המטרה: סחיטת הביצועים המקסימליים מהמודל הקיים לאחר שטייבנו את הנתונים.

• משימה 2.1 - סלקציית פיצורים (Feature Selection)

- שימוש בשיטות כמו RFE (Recursive Feature Elimination) כדי לזרוק משתנים "רועשים" שלא תורמים לדיוק ורק מבלבלים את המודל. המטרה: להישאר עם סט מזוקק של 10-15 משתנים משפיעים.

• משימה 2.2 - כיול היפר-פרמטרים (Grid Search)

- הרצת סריקה ממוחשבת למציאת ההגדרות הטובות ביותר למודל: עומק העץ המקסימלי, מספר העצים ביער, וכמות הדגימות המינימלית לעלה. פעולה זו נועדה למנוע Overfitting ולשפר את ההכללה.



8.3. שלב 3: ניתוחי מעטפת וסימולציות (העשרת הדוח)

המטרה: יצירת נפח עבודה נוסף המדגים את יכולות המערכת מעבר למספר היבש של הדיוק.

• משימה 3.1 - ניתוח רגישות (Sensitivity Analysis)

- מיפוי הגורמים המשפיעים: יצירת גרפים המראים אילו פרמטרים הכי משפיעים על ה- EC (האם זה המים? הטמפרטורה? הדשן?). זהו פרק חשוב בדיון המסכם.

• משימה 3.2 - בדיקת תרחישי קיצון (Stress Testing)

- הרצת המודל על נתונים סינתטיים קיצוניים (למשל: "מה קורה לחיזוי אם יש שרב קיצוני וההשקיה מפסיקה?"). זה יאפשר לנו להראות שהמודל מתנהג בצורה "הגיונית" גם במצבים שלא קרו בפועל.

• משימה 3.3 - השוואה למודלים נאיביים:

- הוכחת עליונות המודל על ידי השוואה לחיזוי טיפש (כמו "הערך מחר יהיה זהה לערך היום"). זה יוסיף נפח סטטיסטי לעבודה.

8.4. שלב 4: סיכום וכתובה

- **משימה 4.1:** כתיבת ספר הפרויקט, הפקת גרפים מסכמים ומצגת.