

הטכניון - מכון טכנולוגי לישראל

הפקולטה להנדסת חשמל

מעבדת SAMPL

**דו"ח סיכום פרויקט**

Sugar Project

**מבצעים:**

Opher Bar Nathan	302188628	עפר בר נתן
------------------	-----------	------------

Nadav Eitan	302957626	נדב איתן
-------------	-----------	----------

**מנחה: אורן סולומון**

**סמסטר רישום: חורף תשע"ח**

**תאריך הגשה: אוגוסט 2018**

3	.....	Abstract – תקציר
4	.....	הקדמה
4	.....	מטרות הפרויקט
5	.....	פרק א' – מבוא תיארוטי
5	.....	RNN - Recurrent Neural Network
5	.....	LSTM - Long Short-Term Memory
6	.....	הנחות ושיטת ניבוא
6	.....	הנתונים
8	.....	פרק ב' – תוצאות
8	.....	מבוא
8	.....	תוצאות
23	.....	מסקנות
23	.....	כיווני מחקר / רעיונות עתידיים
23	.....	References

סכרת היא מחלה מטבולית המתאפיינת בריכוז גבוה של גלוקוז בדם ולעיתים גם בשתן. המחלה הוכרזה לאחרונה כמגיפה עולמית על ידי ארגון הבריאות הלאומי, עם מאות מיליוני חולים ברחבי העולם, ועוד רבים במצב טרום סכרת. חולי סכרת צריכים לעקוב אחר רמות הסוכר בדם שלהם כל הזמן, וזאת בגלל הנטייה של היפרגליקמיה והיפוגליקמיה המתארים רמות גבוהות מאוד ונמוכות מאוד של גלוקוז בדם. המעקב נעשה על ידי בדיקת דם עצמית באמצעות מד-סוכר (גלוקומטר). תדירות הבדיקה משתנה בהתאם לסוג הסוכרת ואופן הטיפול הנבחר ויכולה אף להגיע ל-6–8 פעמים ביום. בשנים האחרונות נוספו גם מדי-סוכר רציפים המאפשרים, במגבלות שונות, מעקב רציף לאורך כל היום והלילה. אחת הדרכים המבטיחות למדוד רמות גלוקוז היא עם מערכת Freestyle Libre, אשר מחוברת עם מחט לחלק האחורי של יד החולה. בפרויקט זה נפתח (RNN) Recurrent neural network על מנת לחזות רמות של גלוקוז בדם, בהתבסס על הנתונים שנאספו מ-Freestyle Libre.

Diabetes is a group of metabolic disorders which results in high blood sugar levels over a prolonged period. Diabetes was recently declared as a worldwide epidemic by the National Health Organization, with hundreds of millions of patients worldwide, and many more in a pre-diabetic state. Diabetes patients need to monitor their blood glucose levels constantly since they are prone to high (hyperglycemia) and low (hypoglycemia) spikes in their blood glucose. Monitoring is possible with self-testing and different kinds of glucometers. The frequency of the testing varies depending on the type of diabetes and treatment chosen and can even reach 6-8 times a day. In recent years, continuous glucose meters have been added, allowing for continuous monitoring throughout the day and night. One of the promising ways to measure glucose levels is with the Freestyle Libre System, which is attached with a needle to the back of the patient's hand. In this project we develop a Recurrent neural network (RNN) to predict blood glucose levels, based on data collected from Freestyle Libre.

פרויקט זה מציע ניבוי הסתברותי עבור מדידות סוכר עתידיות. בעזרת ניבוי עתידי שכזה, כל חולה סכרתי יכול לשפר משמעותית את שליטתו ברמות הסוכר בדם, כאשר המטרה היא להמצא בטווח נורמלי של סוכר בדם. בין אם זה להזריק יותר אינסולין כדי למתן עליה עתידית, ובין אם זה לאכול או לשתות משהו מתוק כדי למנוע ירידת סוכר עתידית. כדי לנבא, בנינו מערכת לומדת אשר מחזיקה בנתונים עבור חולה מסוים ובעזרתו נותנת תוצאות. הקוד בנוי בצורה גנרית כך שכל חולה יוכל להכניס את היסטורית קריאות הסוכר שלו על מנת לקבל תוצאות מותאמות אישית.

## מטרות הפרויקט

הצגת תוצאות עם הסתברות גבוהה וניבוי מדויק של רמות סוכר, כמו כן, הרחבת הידע התיאורטי והפרקטי ברשתות נוירונים, בפרט בRNN ו LSTM שנרחיב עליהם בהמשך ובנוסף, לימוד שפת פייתון.

מטרת העל היא אפליקציה שעובדת בזמן אמת ומציגה לכל חולה ניבוי עתידי ומדויק לרמות הסוכר שלו.

המחשה ויזואלית / הצעת הגשה:

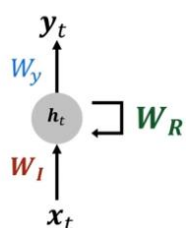


כאן רואים מדידות סוכר עד הנקודה השחורה ומשם ניבוי של מגמת העתיד הנראית כעליה עם ביטחון של 87% בניבוי מדויק.

## RNN - RECURRENT NEURAL NETWORK

נתחיל מאיפה נולד הצורך של ארכיטקטורת RNN: ברשתות feed-forward רגילות, הקלט הוא באורך קבוע והקלט בלתי תלוי אחד בשכנו, אך במקרים רבים ובמיוחד בבעיות של קלט שמגיע בצורת רצף, אורך הקלט יכול מאוד להשתנות וגם יכולה להיות תלות חזקה בין דגימה אחת לשניה, בטווח הקצר וגם בטווח הארוך. תחת ההנחות האלה, RNN היא סוג של רשת נוירונים שבה החיבורים בין הצמתים יוצרים גרף מכון ומעגלי, שלא כמו רשתות נוירונים אחרות בהן קשרי הצמתים אינם יוצרים מעגלים. צורה זו מאפשרת התנהגות דינמית ברצף בזמן ובכך RNN יכול להשתמש במצב הפנימי שלו כדי לעבד רצפים של קלטים. פעולה זו הופכת את הרשת לשימה עבור משימות המורכבות מרצף של קלטים כגון זיהוי כתב יד, עיבוד שפה טבעית (NLP).

## Recurrent neuron



$$h^{(t)} = g_h(W_I x^{(t)} + W_R h^{(t-1)} + b_h)$$

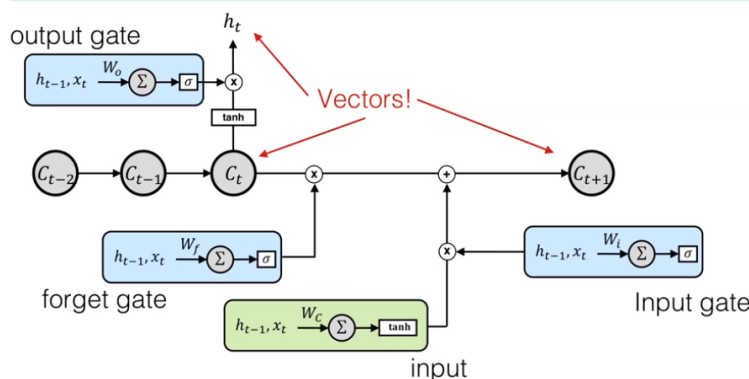
$$y^{(t)} = g_y(W_y h^{(t)} + b_y)$$

## LSTM - LONG SHORT-TERM MEMORY

יחידות LSTM הם יחידות בניין עבור שכבות RNN. כאשר RNN מורכב מיחידות LSTM ניתן לקרוא לו פשוט LSTM. יחידה בסיסית מורכבת מתא, שער לקלט, שער לפלט ושער שיכחה. התא אחראי על לזכור ערכים על פני מרווחי זמן שרירותיים וכל אחד משלושת השערים יכול להיחשב כנוירון, כלומר, כל אחד מחשב את פונקציית ההפעלה לסכום משקולות. באופן אינטואיטיבי, הם יכולים להחשב הרגולטורים של זרימת הערכים שעוברים בקשרים בין תאי LSTM ומכאן השם שער. ב-LSTM ככל ש"מתקדם הזמן" הפעילות של כל תא זכרון היא לרשת את הפעילות מהתא הקודם בזמן. אבל פעילות זו קצת מנוגת, לכן נרצה גם "לשכוח" מהתא הקודם, גם "להוסיף" זכרון חדש וגם "לקבל" מזכרון קיים.

ארכיטקטורת LSTM מותאמת היטב לסיווג, עיבוד וניבוא של רצף תצפיות במרווח זמן נתון וזאת כי בנייתה הייתה כך שכאשר רשת LSTM מאומנת אז בעית העלמות הגרדיאנט (תוצר מתמטי) תעלם וכמו כן נוצרת שרשרת של השפעות תלות בין רצפי הקלט. באופן אינטואיטיבי, שער הכניסה שולט באיזה מידה זורם ערך חדש לתא, שער השיכחה שולט באיזה מידה נשכח הערך הקודם מהתא ושער הפלט שולט באיזה מידה נקרא מתא הזכרון וזה מהווה את הליבה של רשת LSTM.

## LSTM



אחת מהסיבות שסכרת היא מחלה שקשה לניהול היא קושי ההחלטה של כמה אינסולין להזריק עבור כל ארוחה וכמה אינסולין צריך להזריק באופן קבוע לגוף. מערכת לומדת היא אמצעי לזיהוי מגמות שאדם לא מסוגל להבחין. יכולת זו יכולה לתת תוצאות מעניינות ומועילות לניהול המחלה. אנו מניחים כמה הנחות כדי להקל על המערכת הלומדת כדי שתוכל לנבא תוצאות יותר מדויקות. עיקרן מתבססות על מחזוריות אורח החיים של בנאדם.

מחזוריות היום של אדם מתבטאת במחזוריות של תזמוני ארוחות, פעילות ספורטיבית, שעות פעילות שינה ועוד. לדוגמא, אם אדם קם ב7 בבוקר, ביצע פעילות ספורטיבית ואכל ארוחות מסודרות לכל אורך היום אז נרצה שהמערכת תוכל לתת לו ניבוי מאוד מדויק על סמך היכרות המערכת עם ההיסטוריה. לעומת זאת, אם הוא מתנהג באקראיות מסוימת נרצה שהמערכת עדיין תוכל להמליץ לו על סמך ניבוי מסוים. בתקווה שעם הזמן המערכת גם תוכל ללמוד על אותה האקראיות. בסופו של דבר לכולם יש דפוסים חוזרים כלשהם.

שיטת הניבוא שעליה אנו נשענים מתבססת על כמה **משתנים**:

1. כמות זמן החיזוי העתידי – זהו מדד שקובע כמה זמן אנו רוצים להציג החל מרגע מדידה. לדוגמא אם מדדנו ב14:30, נוכל להציג ניבוי **לשעה** עד 15:30 ובכך המשתמש יוכל להחליט איך עליו להתנהג מבחינת ניהול.
2. כמות זמן עליו מסתמכים בהיסטורית המדידות – לדוגמא אם משתמש מודד סוכר ב14:30, נרצה שהמערכת תנבא על סמך **שעתיים** קודמות. ההיגיון עליו אנחנו מתבססים הוא שהסוכר מושפע מהזמן האחרון.
3. מספר שעות שגורמות למדידות יום להיות לא רלוונטיות - ישנם ימים בהם אין מספיק נתונים מהמכשיר מסיבות כמו החלפה דו שבועית של סנסור או זכרון מוגבל של המכשיר. עבור ימים בהם חסרים **שלוש שעות** נרצה שלא יפגע בתהליך הלימוד ולכן נרצה להסיר את כלל מדידות היום כליל.

## הנתונים

הנתונים המגיעים ממכשיר הליברה ניתנים להעברה למחשב מהמכשיר, כך שרמות הסוכר מיובאות בצורה

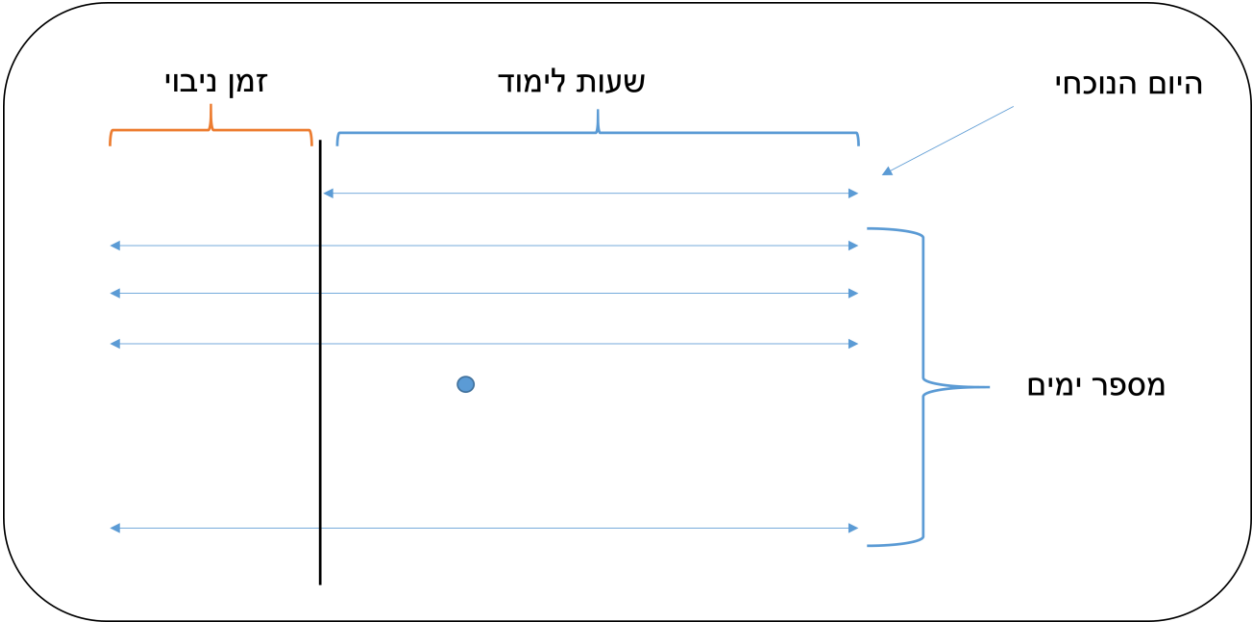
2016/11/18 20:29	0	81
2016/11/18 20:44	0	81
2016/11/18 20:59	0	84
2016/11/18 21:14	0	90
2016/11/18 21:29	0	99

שהם תאריך ושעת מדידה לעומת רמת סוכר.

הכנת המידע בצורה מתאימה למערכת לומדת כללה כמה שלבים חשובים:

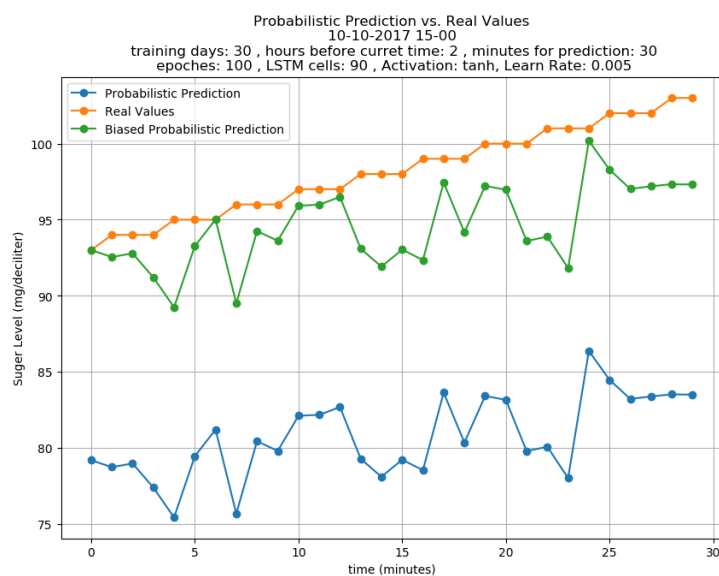
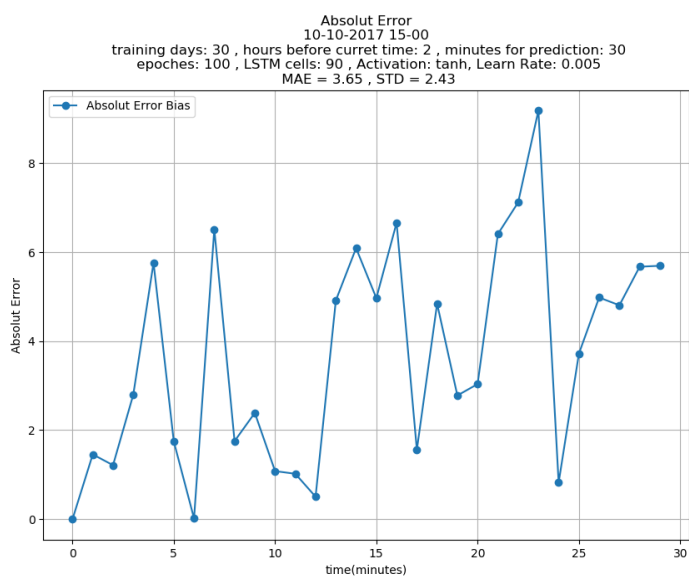
1. כדי שתוצאות החיזוי יהיו ברזולוציה של דקה בודדת הרחבנו, על ידי אינטרפולציה לינארית, את המדידות הנתונות לרזולוציה של כל דקה.
2. נירמול - "הלבנה" של כל המדידות.
3. סידור (מטריצי) בצורה של ימים בשביל שתתקבלנה תוצאות המבטאות את אותה מחזוריות עליה אנחנו מסתמכים.

תיאור גרפי של צורת המידע עור אימון הרשת:



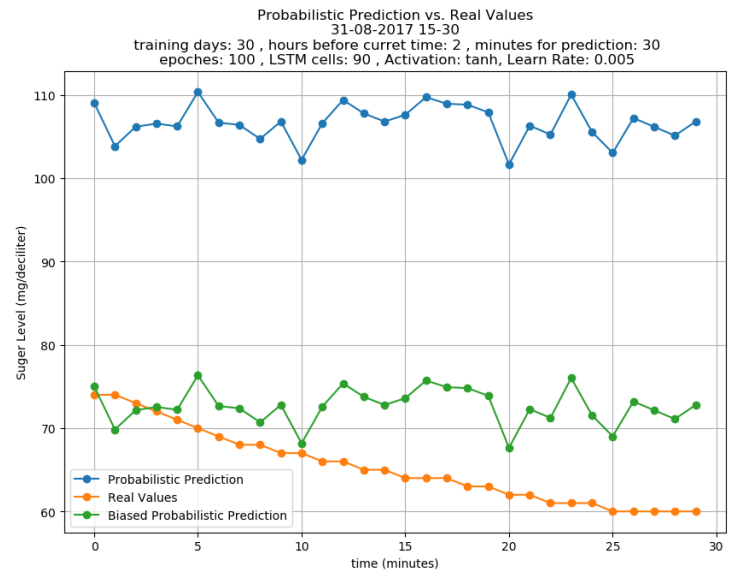
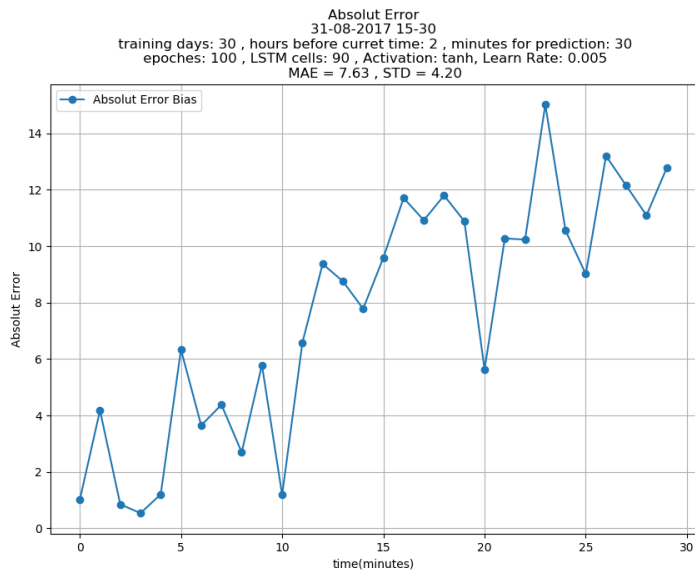
- מבחינת חולי הסכרת יש משמעות למרחק חיזוי התוצאות מהערך האמיתי (ולא המרחק הריבועי לדוגמא) ולכן בחרנו להציג שגיאת MAE (Mean Absolute Error) כדי לבחון את טיב התוצאות
- נבדוק את תוצאות החיזוי עבור ערכים שונים של פרמטרי האימון ופרמטרי רשת הניורונים:
- פרמטרי האימון: מספר ימי האימון, שעות לימוד, משך זמן הניבוי
- פרמטרי רשת הניורונים: מספר תאי ה LSTM, קצב הלימוד, פונקציית אקטיבציה, איפוצ'ס
- בחרנו שלושה זמנים בהם היו עליית סוכר, ירידת סוכר וללא מגמה

### מגמת עלייה – נראה שיש מגמה תואמת עבור ערכי הניבוי.

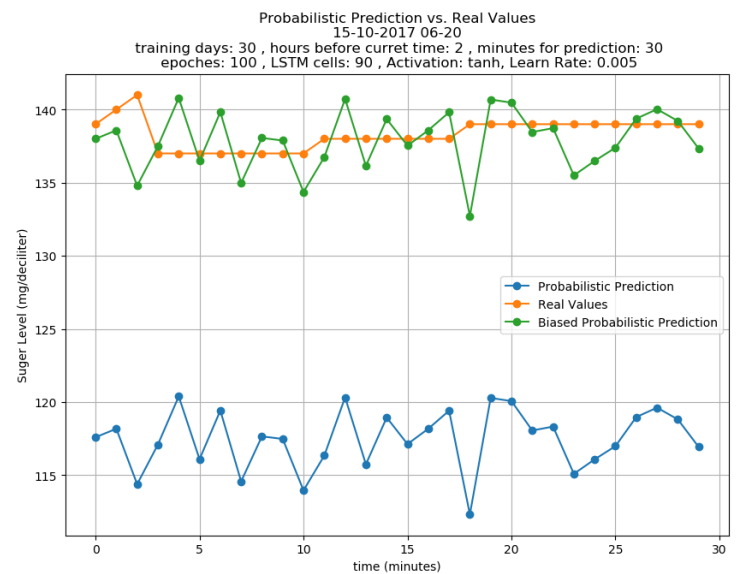
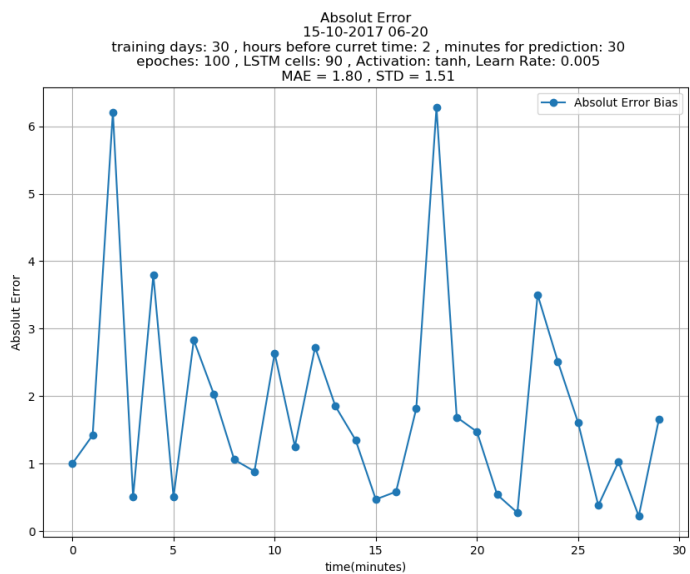




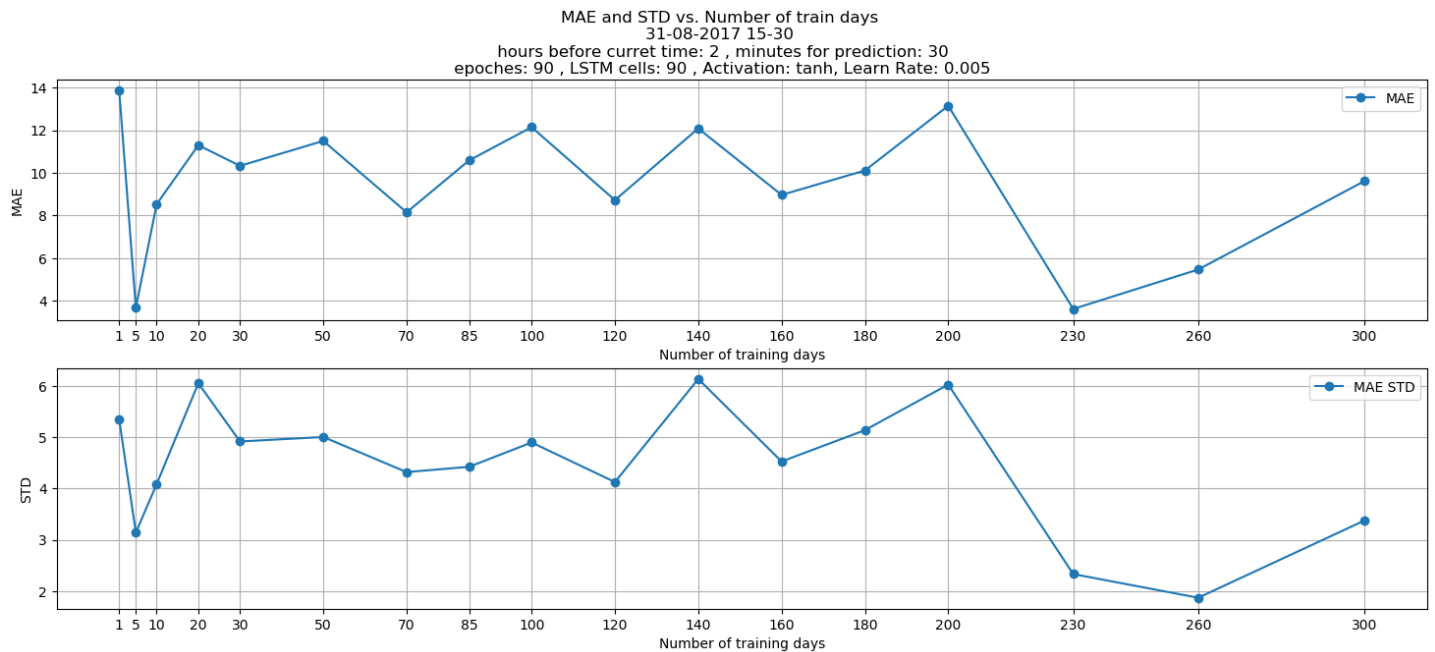
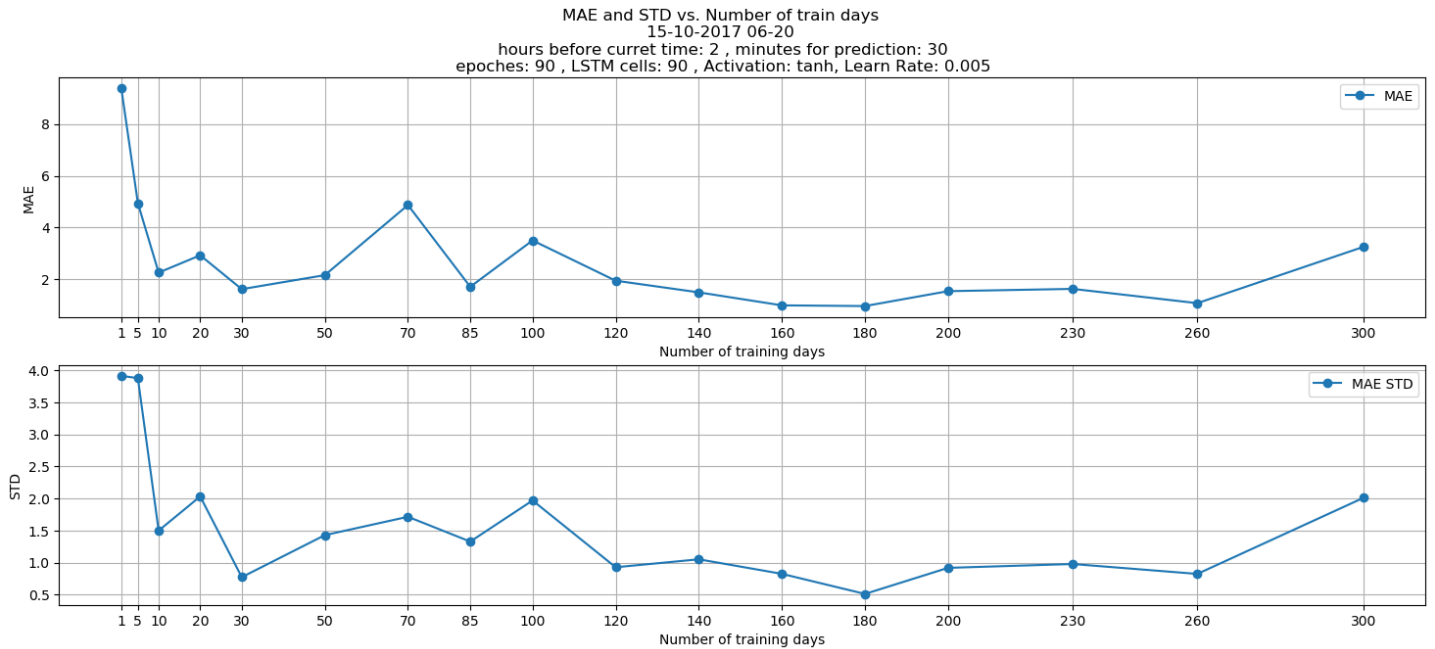
## מגמת ירידה – נראת מגמה של ערך קבוע – מה שאינו מתאים לדגימות האמיתיות



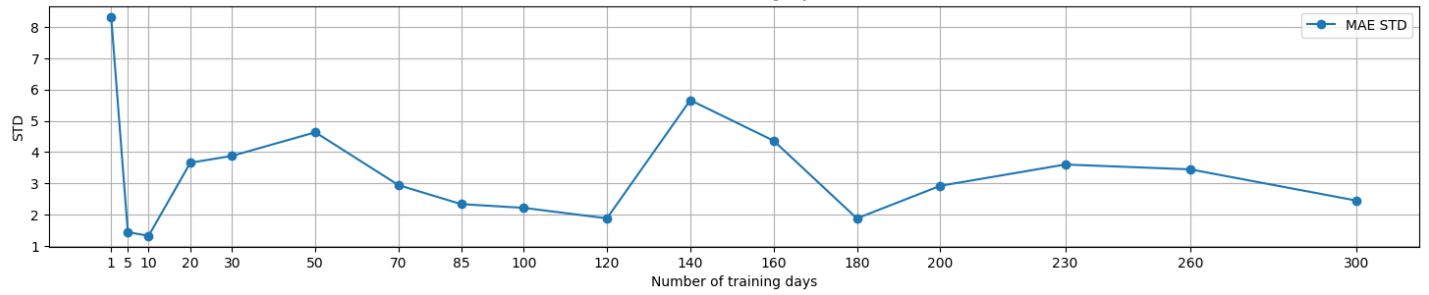
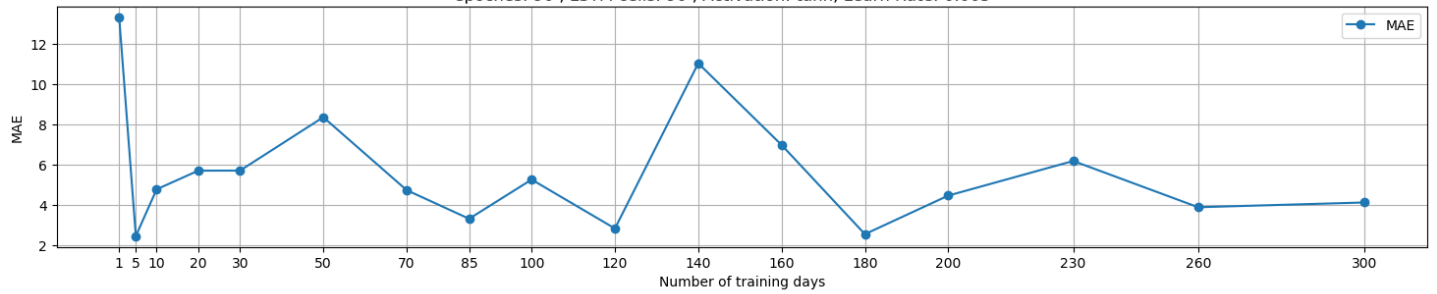
## ללא מגמה – חיזוי מתאים



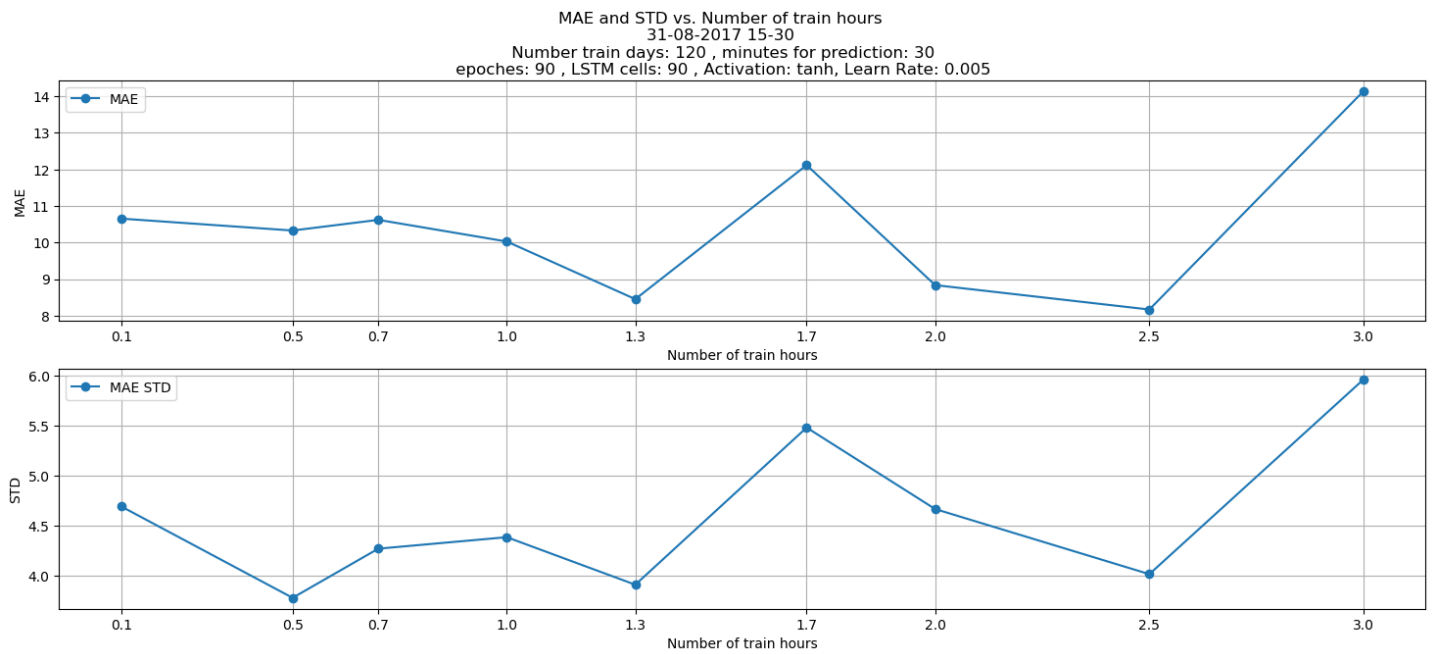
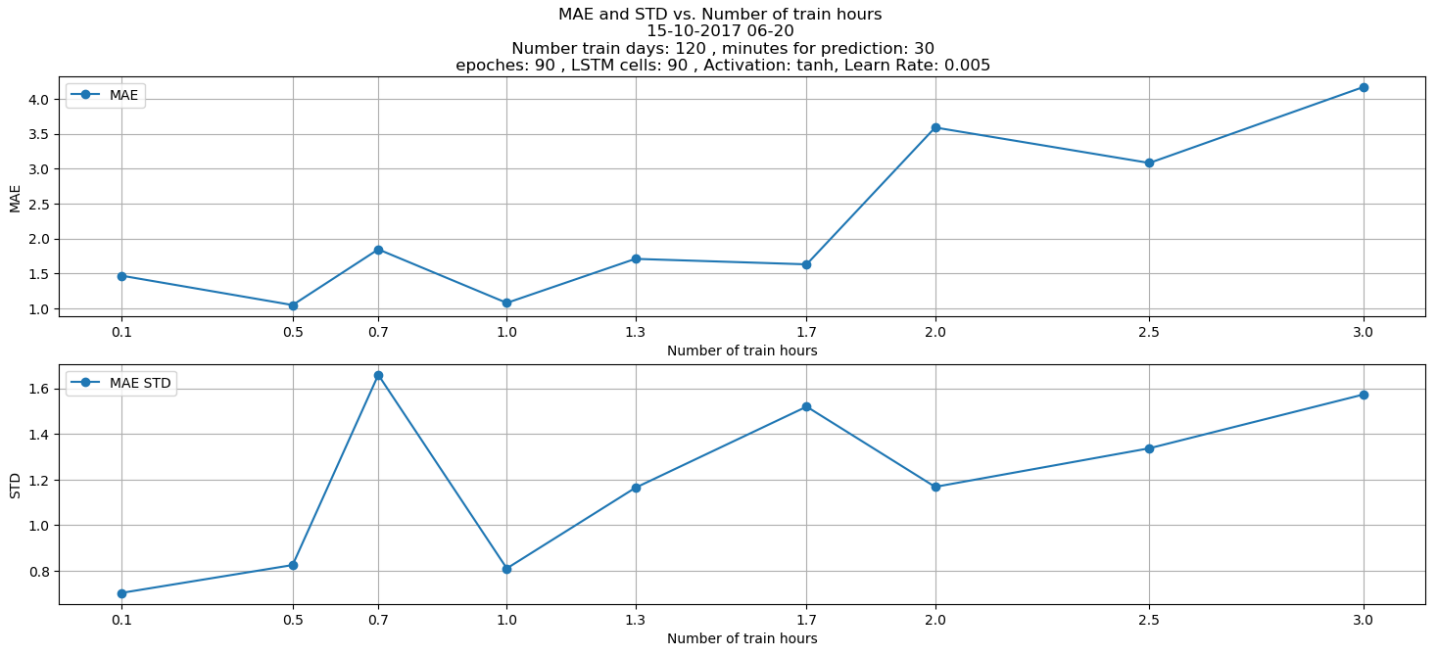
## מספר ימי האימון – לא ראינו מגמות משותפות



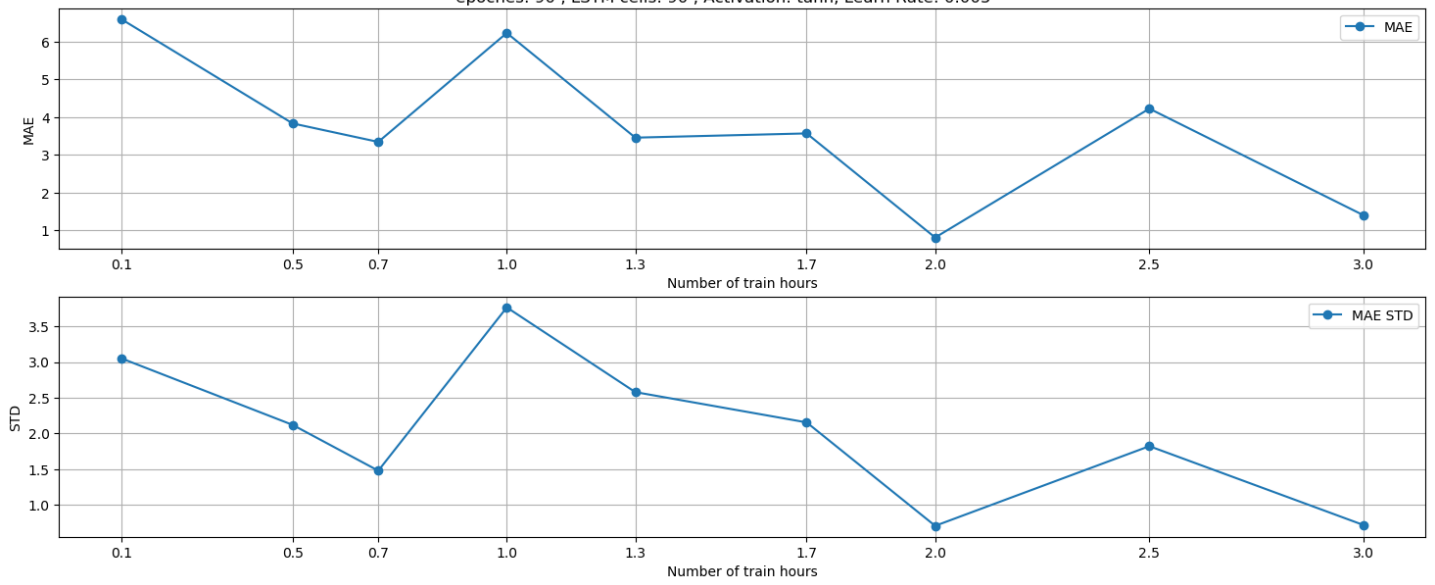
MAE and STD vs. Number of train days  
10-10-2017 15:00  
hours before current time: 2, minutes for prediction: 30  
epochs: 90, LSTM cells: 90, Activation: tanh, Learn Rate: 0.005



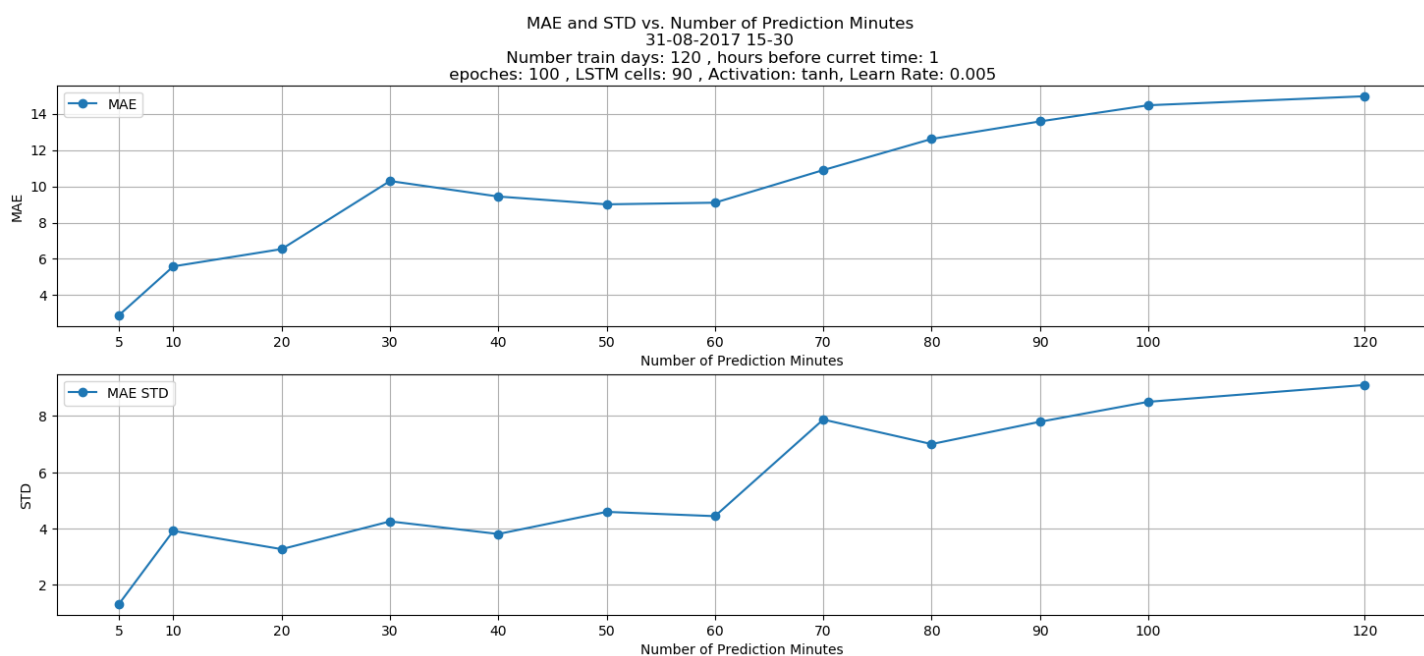
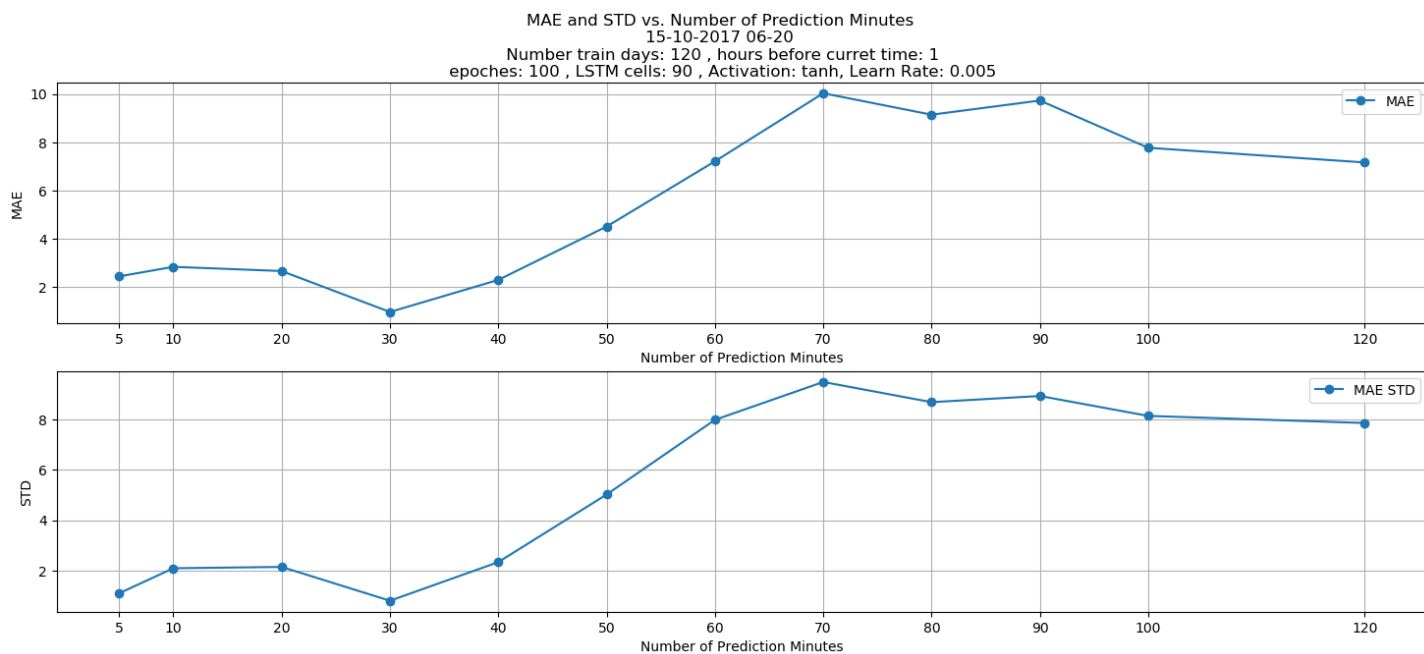
## שעות לימוד - לא ראינו מגמות משותפות



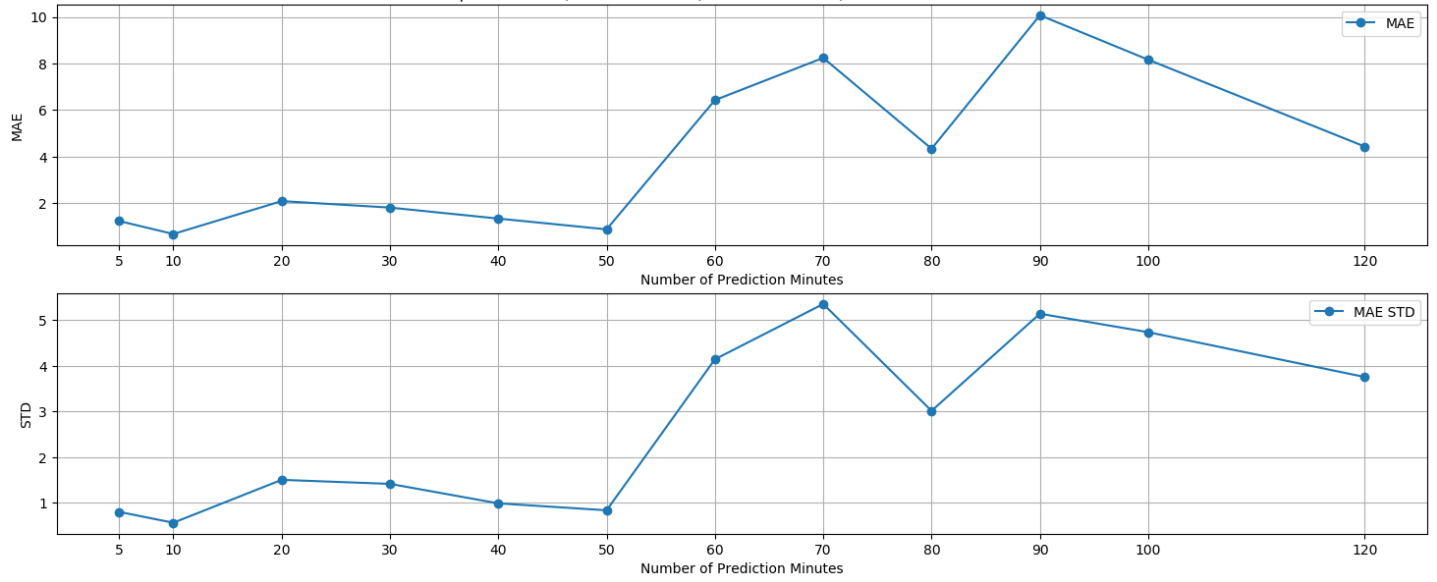
MAE and STD vs. Number of train hours  
10-10-2017 15-00  
Number train days: 120 , minutes for prediction: 30  
epoches: 90 , LSTM cells: 90 , Activation: tanh, Learn Rate: 0.005



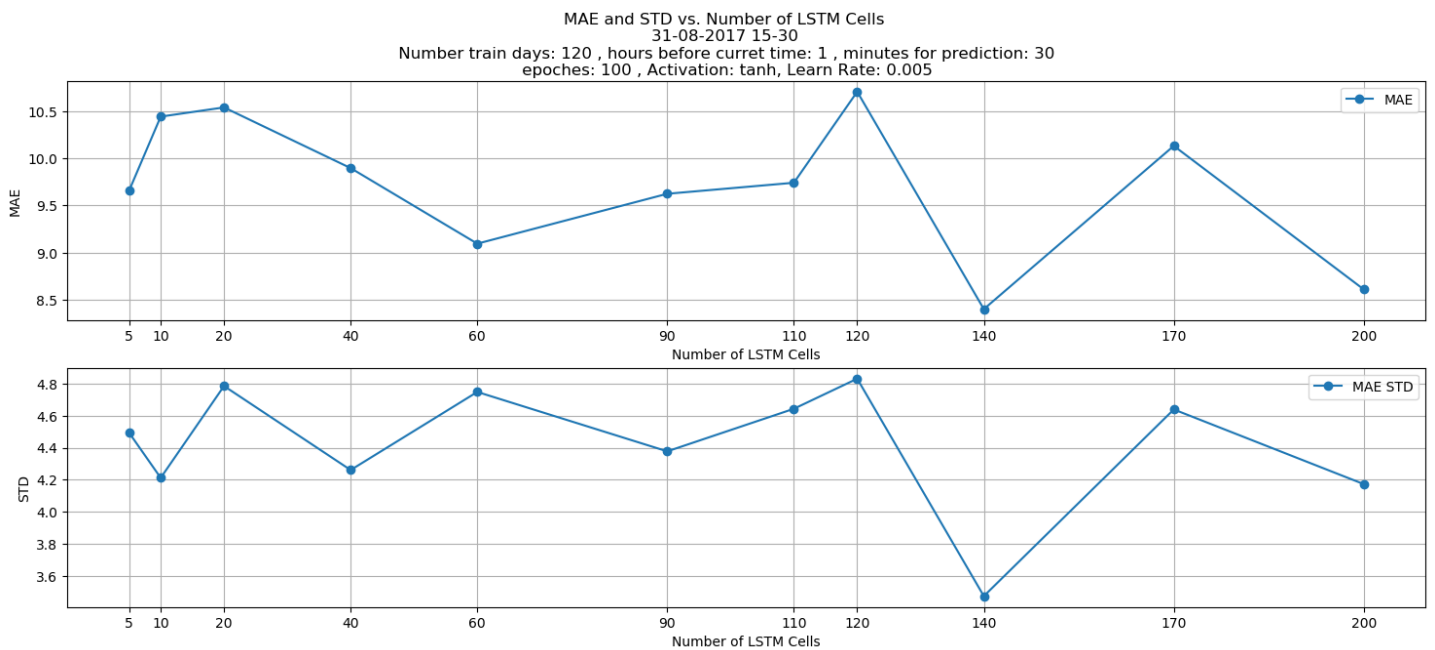
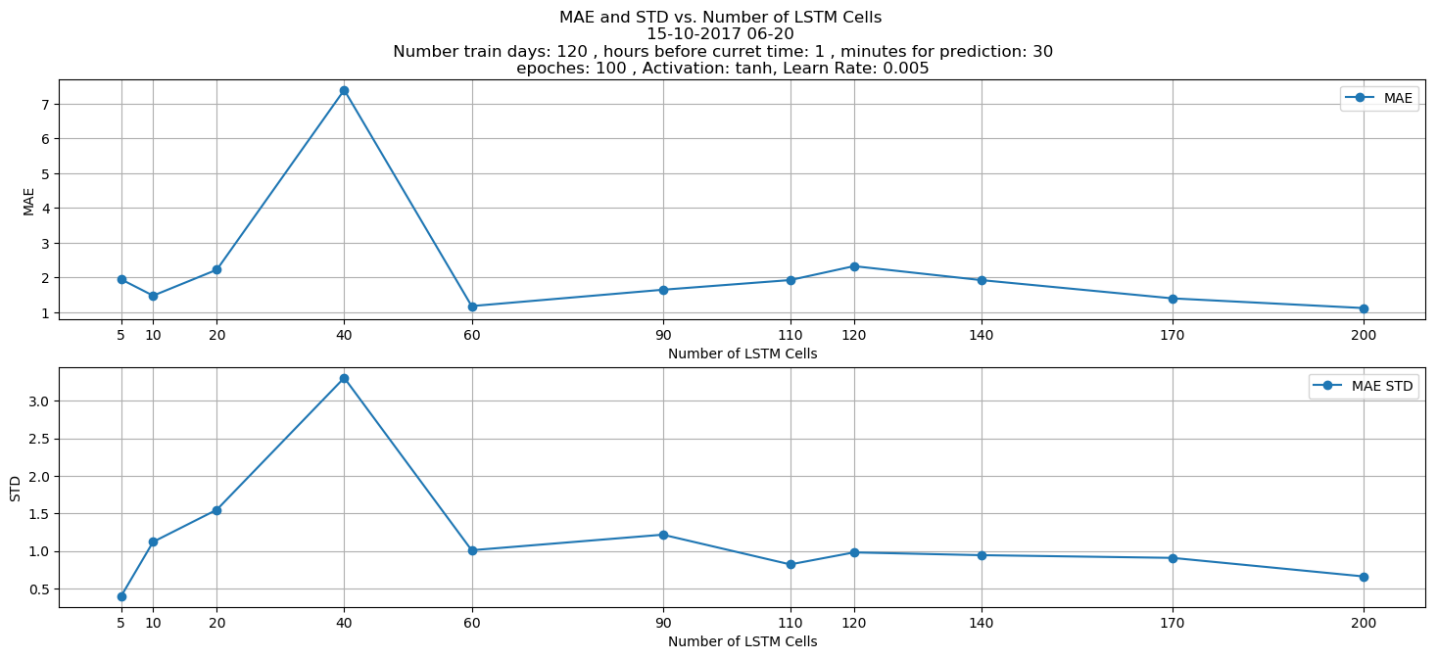
משך זמן הניבוי – בהתאם לאינטואיציה, רואים בבירור שככל שמשך זמן הניבוי גדול יותר, השגיאה עולה. אין ערך אחד שנראה יותר דומיננטי מהשאר. ניכר מינימום מקומי בזמן שאינו היה צפוי (בדקה 30 בגרף ללא מגמה ו50 בגרף עליה).



MAE and STD vs. Number of Prediction Minutes  
10-10-2017 15-00  
Number train days: 120 , hours before curret time: 1  
epoches: 100 , LSTM cells: 90 , Activation: tanh, Learn Rate: 0.005

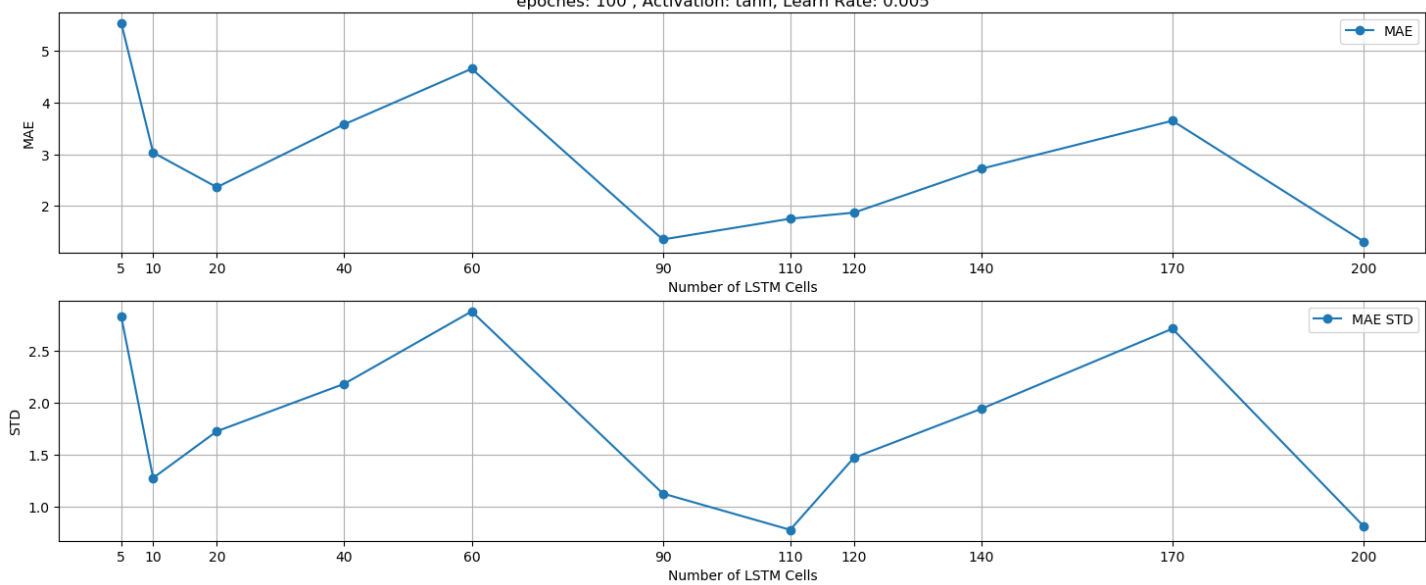


## מספר תאי ה-LSTM - לא ראינו מגמות משותפות

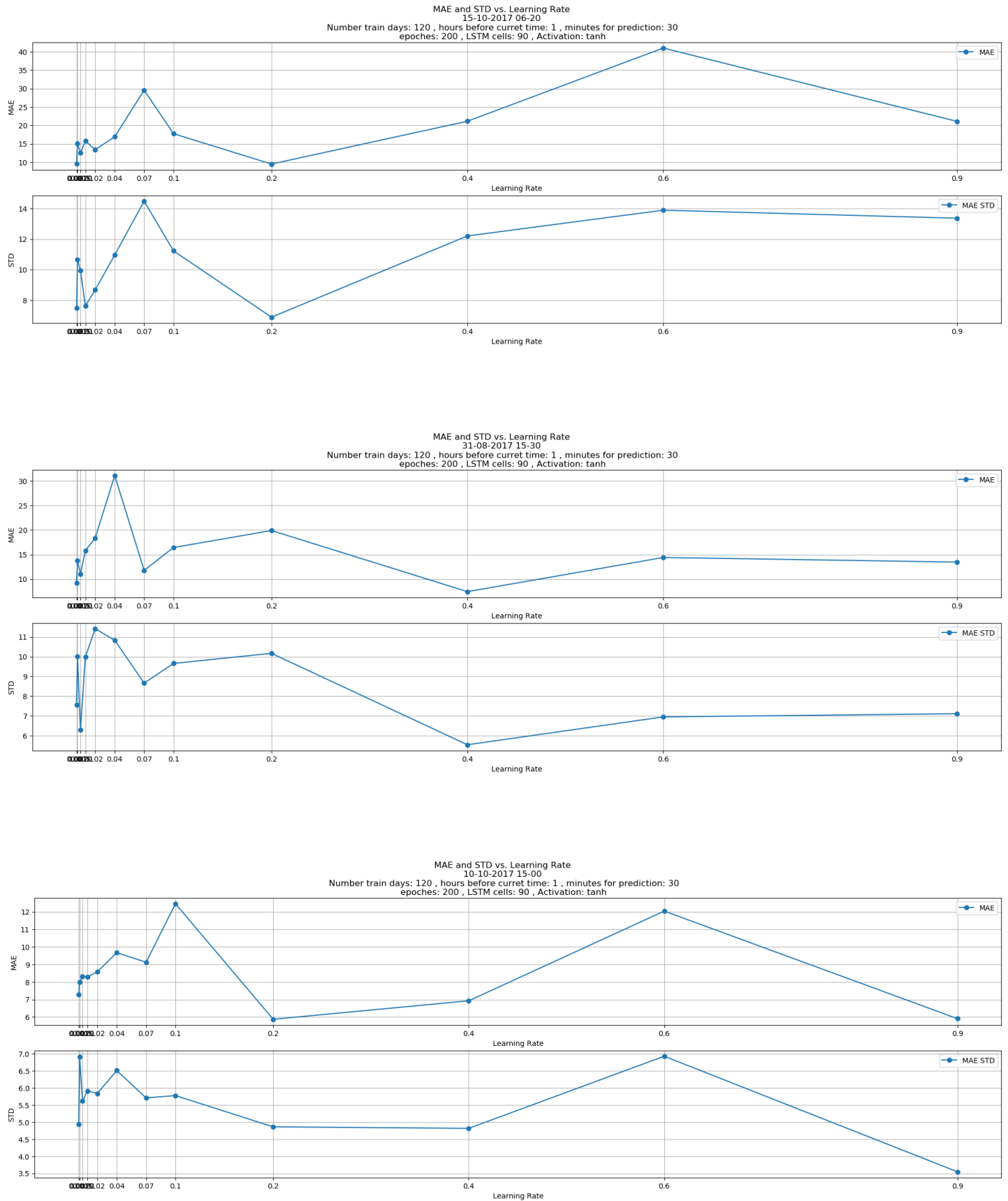




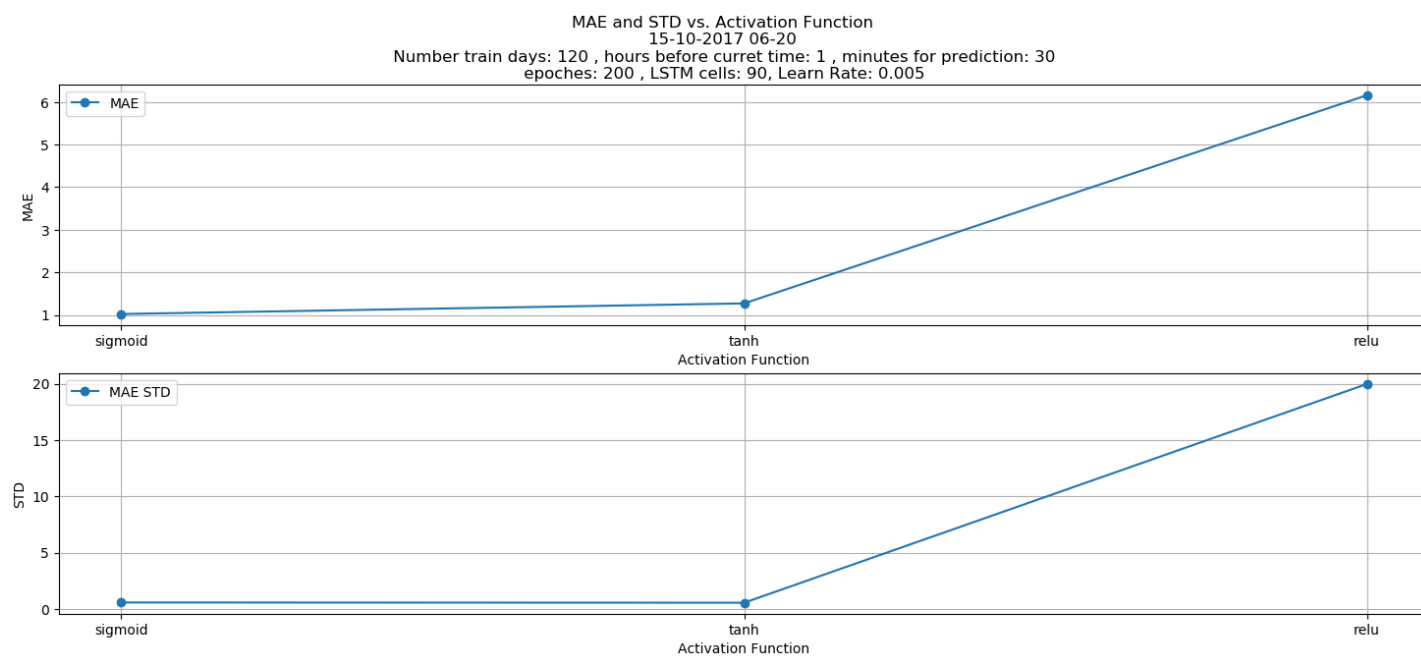
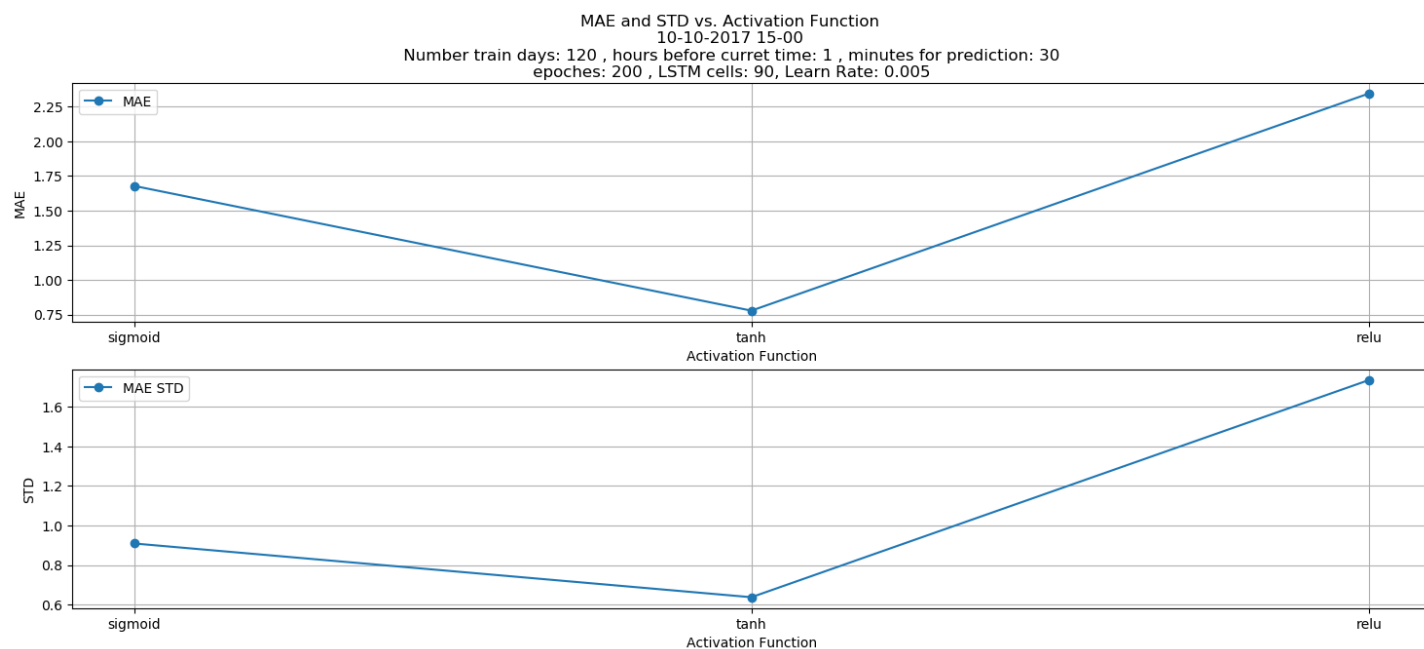
MAE and STD vs. Number of LSTM Cells  
10-10-2017 15-00  
Number train days: 120 , hours before curret time: 1 , minutes for prediction: 30  
epochs: 100 , Activation: tanh, Learn Rate: 0.005



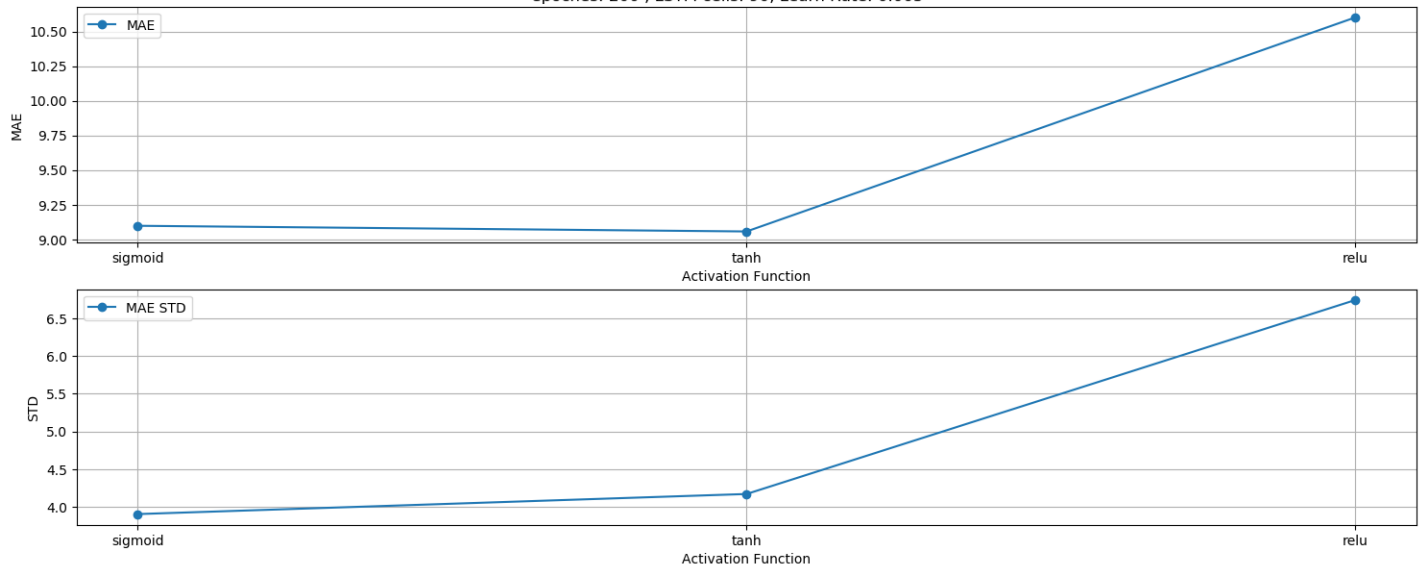
## קצב לימוד – בכל התוצאות רואים מינימום מקומי בערך ביניים.



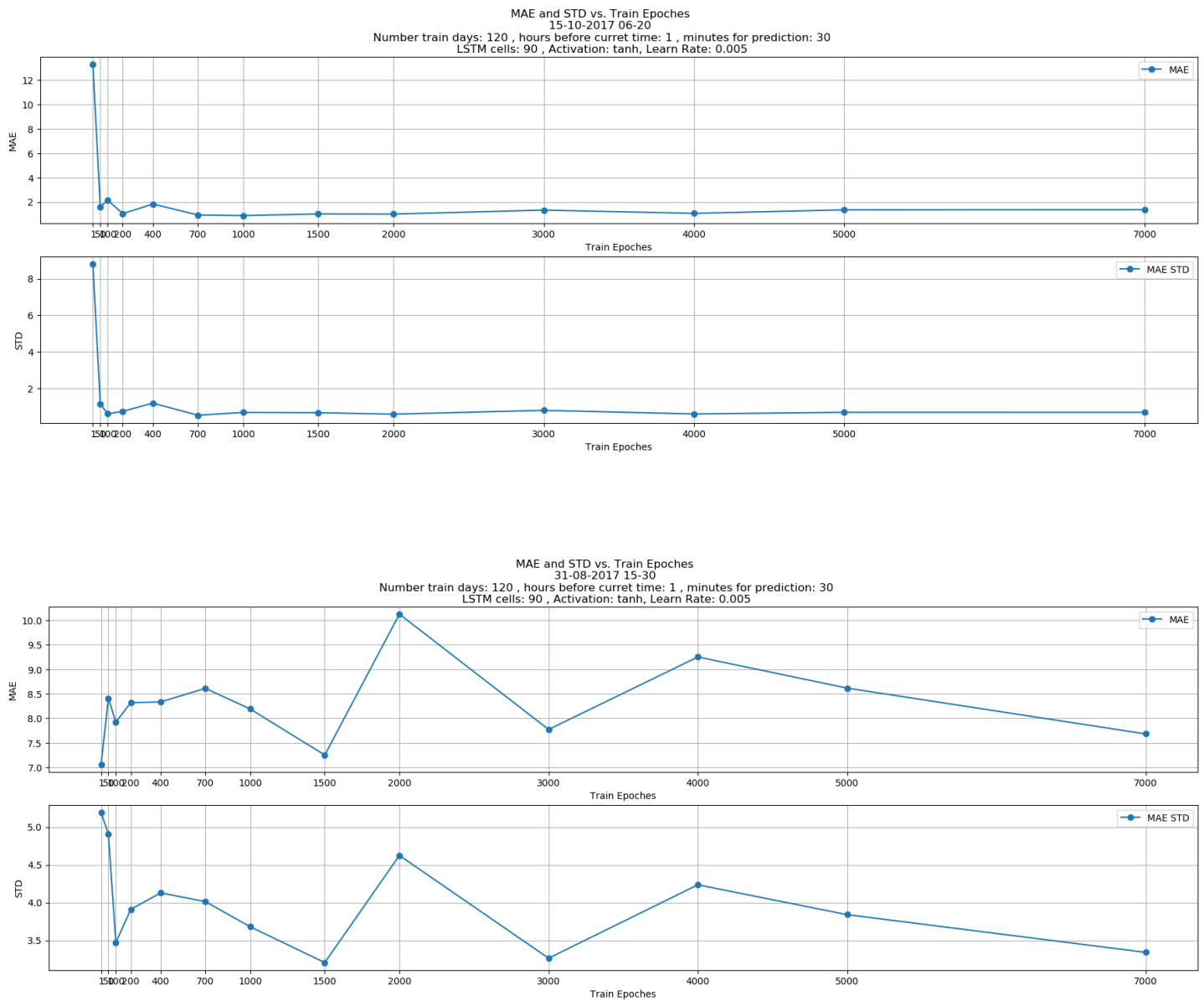
## פונקצית אקטיבציה – בכל התוצאות השגיאה הגבוהה ביותר מתקבלת עבור Relu.



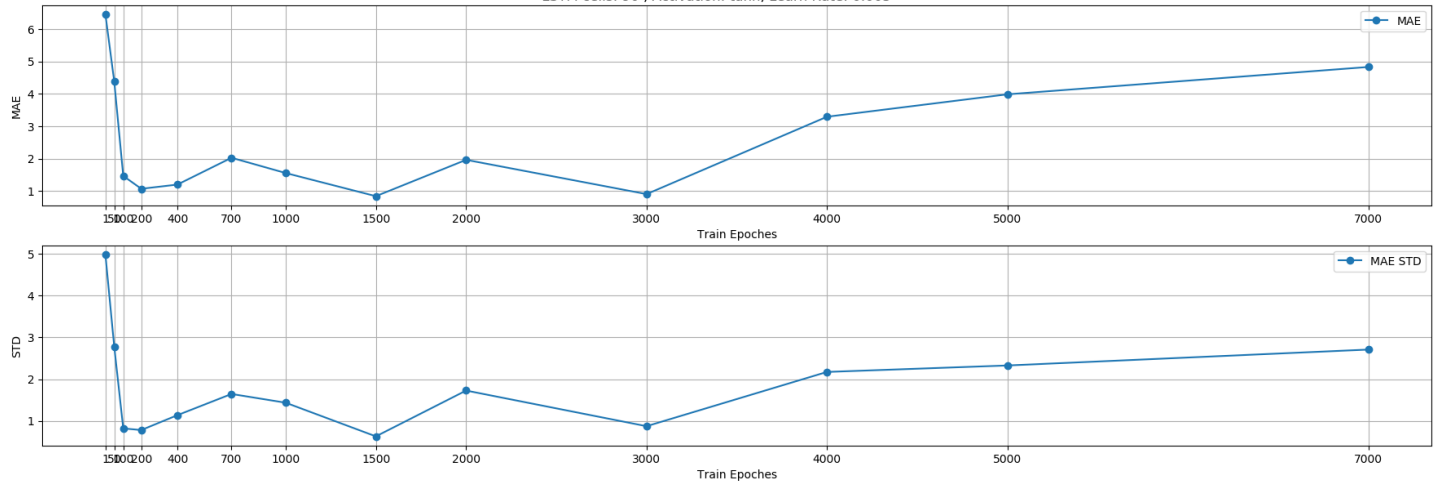
MAE and STD vs. Activation Function  
31-08-2017 15-30  
Number train days: 120 , hours before curret time: 1 , minutes for prediction: 30  
epoches: 200 , LSTM cells: 90, Learn Rate: 0.005



## איפוצ'ס – ניתן לראות שעבור מקרים מסוימים מספר גדול יותר של חזרות אינו משפר את השגיאה.



MAE and STD vs. Train Epoches  
10-10-2017 15-00  
Number train days: 120 , hours before curret time: 1 , minutes for prediction: 30  
LSTM cells: 90 , Activation: tanh, Learn Rate: 0.005



- בחלק מהמקרים, חיזוי המערכת תואם את המגמה של הדגימות האמיתיות
- ערכי השגיאה (Mean Absolute Error) וסטיית התקן (STD) עבור רוב ערכי המשתנים האופטימאליים, הם יחסית קטנים
- בתוצאות המוצגות, חיפשונו ערך אופטימאלי של משתני האימון והרשת, תוך קביעת ערכי המשתנים (היפרפרמטרים) האחרים. קביעת ערכים אלה משפיעה על מציאת ערך המשתנה האופטימאלי. לכן הערכים אשר נמצאו לכל משתנה הם אופטימאליים במובן קביעת ערכי המשתנים האחרים, שילוב שונה יניב ערך אופטימאלי שונה
- למציאת ערכי המשתנים (היפרפרמטרים) האופטימאליים, יש לבדוק שילובים נוספים בין המשתנים השונים
- למען הערכת ביצועי המערכת וקביעת אמינותה דרוש מחקר מקיף יותר ובדיקות נוספות לבדיקות המדגמיות שנעשו, זאת כדי לקבוע אם המערכת אכן מנבאת תוצאות מבטיחות לאפליקציה אמינה

## כיווני מחקר / רעיונות עתידיים

- בחינת שיפורים בתוצאות החיזוי:
- ביצוע אופטימיזציה עבור המשתנים באימון המערכת
- ביצוע חיזוי על בסיס נתונים נוספים כגון כמויות אינסולין, זמני ארוחות
- שימוש ב - RNN מסוג אחר כגון GRU
- ביצוע נירמול שונה עבור הקלט לפני כניסתו למערכת ועבור הפלט לאחר יציאתו מהמערכת
- פיתוח אפליקציה למכשירי smartphone, הלומדת את המשתמש ומבצעת חיזוי מותאם
- ראוי להוסיף הצפנה של המידע

## REFERENCES

- Recurrent Neural Networks - Nervana Systems Archive [https://youtu.be/Ukgii7Yd\\_cU](https://youtu.be/Ukgii7Yd_cU)
- "Modelling Time Series Data with Theano" Hoo Chang Shin – Certified Instructor, NVIDIA Deep Learning Institute [http://www.nvidia.co.kr/content/apac/event/kr/deep-learning-day-2017/dli-2/Modelling\\_Complex\\_Data\\_Sequences.pdf](http://www.nvidia.co.kr/content/apac/event/kr/deep-learning-day-2017/dli-2/Modelling_Complex_Data_Sequences.pdf)
- Keras Documentation <https://keras.io/>