THESIS DEFENSE

CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHM

Student: TRAN HOANG HAI YEN

ID: IELSIU19319

Thesis Advisor: Dr. DAO VU TRUONG SON

24 July, 2024

AGENDA OVERVIEW

01 INTRODUCTION

- Background
- Problem Statement
- Objectives

02 METHODOLOGY

- Literature Review
- Proposed Model: TSK-MBGD

03 SOLUTION DEVELOPMENT

- Model Modification
- Data Preparation
- Train – Test – Validation

04 RESULT & ANALYSIS

- Model Result
- DA-TPA Sensitivity Analysis

05 CONCLUSIONS

- Discussion
- Implication
- Future Research

01

INTRODUCTION

Background

Problem Statement

Objective

INTRODUCTION

BACKGROUND

- Growing population globally → Food security pressure of increasing demand & resource constraints

PROBLEM STATEMENT

- Traditional prediction methods: slow & less accurate
- ML offers efficient & precise solutions, but challenges remain due to complex factors

OBJECTIVE

- Enhance crop yield prediction accuracy using TSK-MBGD algorithm
- Provide insights for farmers, businesses, and policymakers to support sustainable agriculture

02

METHODOLOGY

Literature Review

Proposed Model:
Takagi – Sugeno – Kang Fuzzy with Mini-batch Gradient Descent (TSK-MBGD)

LITERATURE REVIEW

Table 1: *Summary on Literature Review Reference*

Articles	Objective	Research Method
Kaike Sa Teles Rocha Alves, Caian Dutra de Jesus, Eduardo Pestana de Aguiar (2024)	▪ Linear Rule Characteristics ▪ Time Series Forecasting	▪ New Takagi–Sugeno–Kang (NTSK) model with various clusters to form the rules ▪ Recursive least squares (RLS) ▪ Weighted recursive least squares (wRLS)
Qiongdan Lou, Zhaohong Deng (2022)	▪ Multi-Label Learning ▪ Correlation Information	▪ Multi-Label Takagi-Sugeno-Kang Fuzzy System (ML-TSK FS)
None Maryum Bibi, Saif Ur Rehman, & None Khalid Mahmood. (2023)	▪ User-Friendly System ▪ Hybrid Approach (Machine Learning + Deep Learning)	▪ Random Forest (RF) ▪ Artificial Neural Network (ANN)
Agarwal, S., & Tarar, S. (2021)	▪ Hybrid Approach (Machine Learning + Deep Learning) ▪ Cost Efficiency	▪ Support Vector Machine (SVM) ▪ Long-Short Term Memory (LSTM) ▪ Recurrent Neural Network (RNN)
Khaki, S., Wang, L., & Archontoulis, S.V. (2020)	▪ Environmental Data ▪ Management Practices	▪ Convolutional Neural Network (CNN) ▪ Recurrent Neural Network (RNN)

PROPOSED MODEL: TSK-MBGD

01 Takagi – Sugeno – Kang Fuzzy System (TSK)

- Captures **complex** relationships
- Considers **multiple** crop & environmental **factors**

02 Mini-batch Gradient Descent (MBGD)

- Handles large dataset effectively
- Updates parameters in **small batches** → Reduces computational load
- Ensures **stable convergence**

ADVANTAGE

Enhanced Accuracy & Scalability

Robust Optimization

Combined Interpretability & Adaptability

03

SOLUTION DEVELOPMENT

Model Modification

Feature Correlation

Data Preparation

Train – Test – Validation

MODEL MODIFICATION

MODEL REFERENCE SOURCE PyTSK developed by YuqiCui, publicly available on Github https://github.com/YuqiCui/pytsk

NOTED **Regression** Problem (Not *Classification* as developed in ref)
+ Added **MBGD** Algorithm
→ Require Modification

Regression Validation through Evaluation Metrics (R^2 score, RMSE, MAE)

```
from pytsk.gradient_descent.antecedent import
AntecedentGMF, antecedent_init_center
from pytsk.gradient_descent.tsk import TSK

# Import models
from callbackregression import EarlyStoppingRMSE
from trainingregression import Wrapper
```

Mini-batch Element for Wrapper in Training Loop

FEATURE CORRELATION

392 Features

- 6 Weather Elements
- 11 Soil Elements
- 1 Planting Time Element

Most Positive Correlations

- Soil Organic Carbon & Organic Carbon Stock & Density: **0.88 – 0.84**

→ Related to the **organic carbon amount** in the **area**

- Min Temperature & Max Temperature & Vapor Pressure: **0.94 – 0.88 – 0.73**

→ Related to the **overall area's temperature** & **vapor speed**

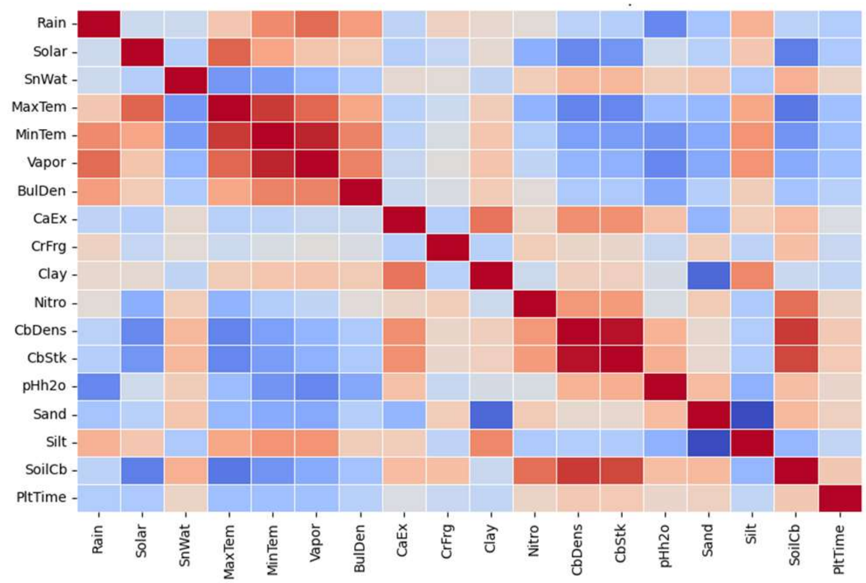


Figure 4.2: *Heatmap of Correlation Between Feature Groups*

Most Negative Correlations

- Sand & Silt: **-0.94**
- Clay & Sand: **-0.80**

→ All 3 **key components** in **Soil Texture & Nutrient Distribution → Mutually exclusive** in **proportions**

- Soil Organic Carbon & Max Temperature/Solar Radiation: **-0.74/-0.71**

→ **High Heat = Increased Decomposition of Carbon**

DATA PREPARATION

1

Data Assessment & Anomalies

- Calculated variance of each feature

$$\sigma^2 = \sum \frac{(x_i - \mu)^2}{N}$$

- Addressed anomalies & missing values **(none)**

2

Handling Missing Data

- Replaced missing data with 0
- Removed columns with constant zeros ***(-13 columns)***

3

Variance Threshold Application

- Established a ***95%*** variance threshold
- Removed high-variance features ***(-19 variables)***

4

Feature & Sample Reduction

- Generated a final dataset with ***360 features***
- Reduced from 25,345 samples to ***1,582 samples*** *(year 2016-2018)*

SOLUTION DEVELOPMENT

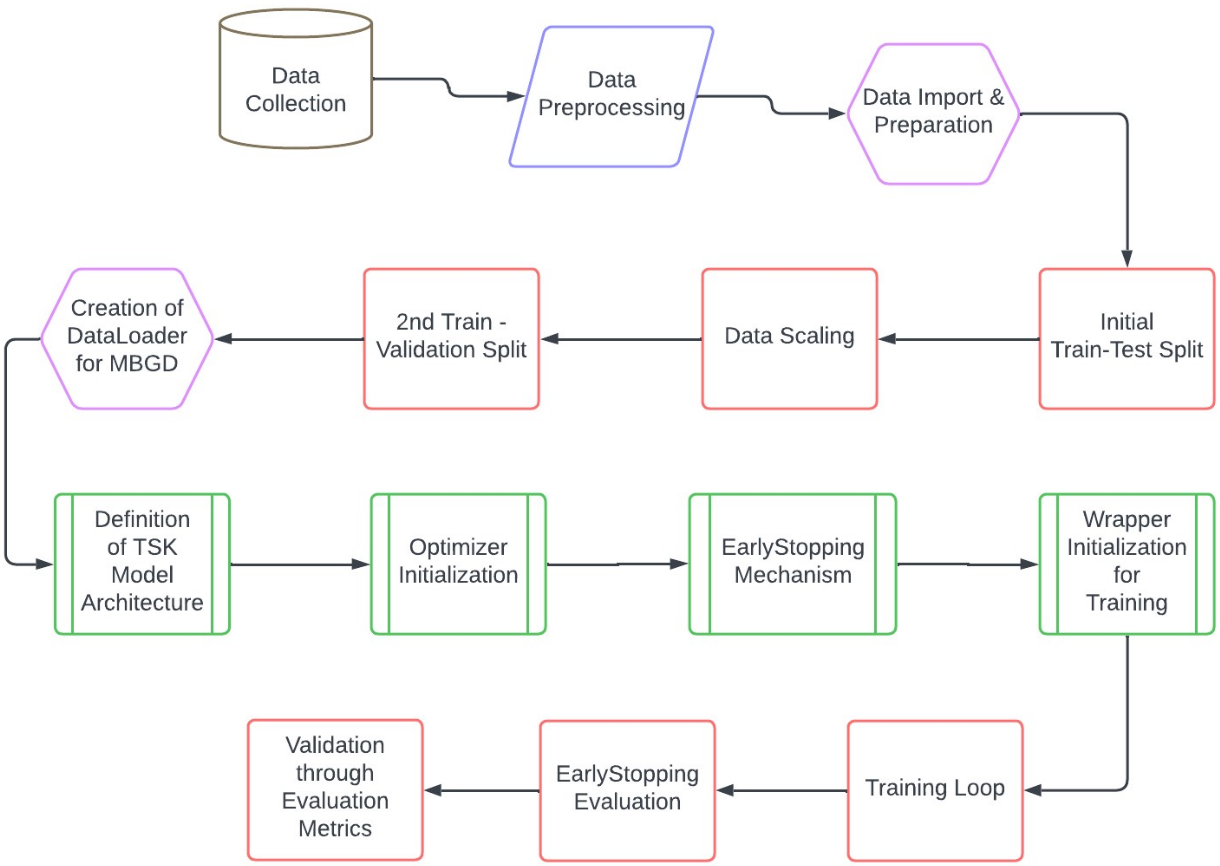


Figure 4.1: *Detailed Process of TSK-MBGD Model for Crop Yield Prediction*

04

RESULT & ANALYSIS

Experimental Result

Model Comparison & Explanation

Delta Analysis - Target Parameter Approach (DA-TPA) Sensitivity Analysis

EXPERIMENTAL RESULT

Fixed Parameters $\alpha\ (learning\ rate) = 0.005,\ K = epochs = 700,\ weight\ decay = 10^{-8},\ patience = 300$

Hyperparameters Batch Size $N_{bs} \in [5, 70]$, step size 2-5

Number of Rules $N_{rule} \in [5, 70]$, step size 2-5

Table 5.1, 5.2, 5.3, 5.4 : *Evaluation Metrics & Running Time at* (N_{rule}, N_{bs})

R^2 Score

#Rules \ Batch Size	15	30	45	60
15	0.7009	0.6758	0.6416	0.6768
30	**0.7211**	0.6820	0.5850	0.6984
45	**0.7580**	**0.7178**	0.6367	0.5217
60	**0.7912**	0.6417	0.5429	0.6179

MAE

#Rules \ Batch Size	15	30	45	60
15	**3.5710**	3.9093	3.8686	3.7781
30	3.7530	3.8953	4.0833	**3.5775**
45	**3.5072**	3.6655	4.0389	4.1791
60	**3.4012**	3.8885	4.4989	4.0878

RMSE

#Rules \ Batch Size	15	30	45	60
15	**4.8925**	5.2592	5.7120	5.2236
30	5.0871	5.4791	5.5258	5.1847
45	**4.7744**	**4.7874**	5.6140	6.4119
60	**4.4331**	5.5789	6.5237	5.8734

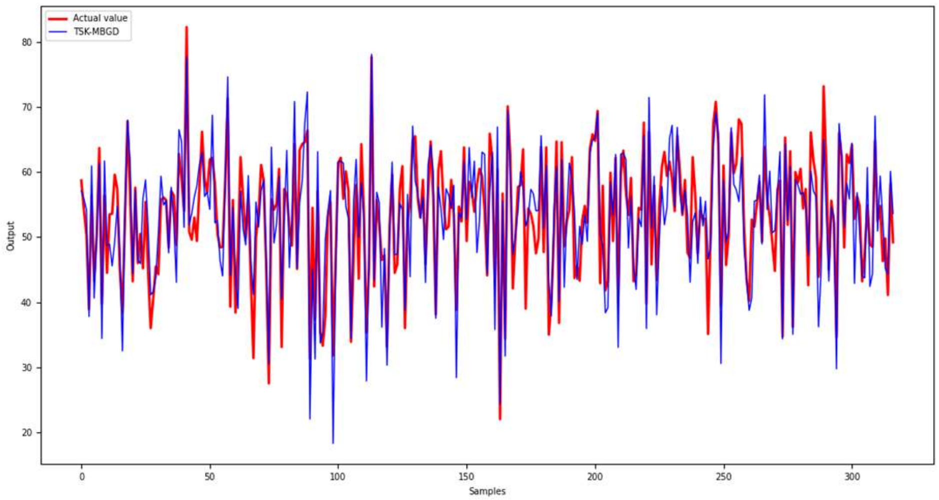
Run Time

#Rules \ Batch Size	15	30	45	60
15	260.43	488.60	331.18	221.09
30	355.20	380.49	368.03	208.60
45	687.65	822.26	412.62	254.12
60	**908.73**	918.38	493.13	548.89

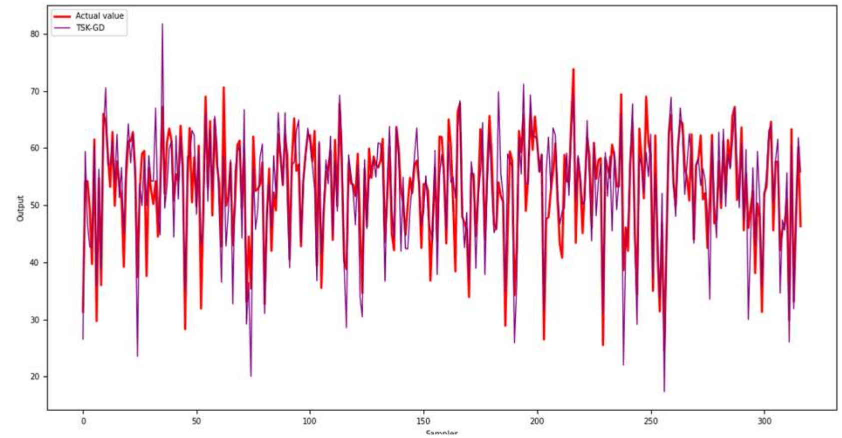
CONCLUSION

- Best performance observed at:
  Batch Size = 15
  Rule = 60 ***(optimal solution)***
- Remarkably Long Running Time (15 minutes)

COMPARISON & EXPLANATION



- **Red**: Actual Value of Dataset
- **Blue**: **TSK-MBGD** (*proposed model*)
- **Purple**: TSK-BGD
- **Green**: NTSK-RLS
- **Black**: NTSK-wRLS



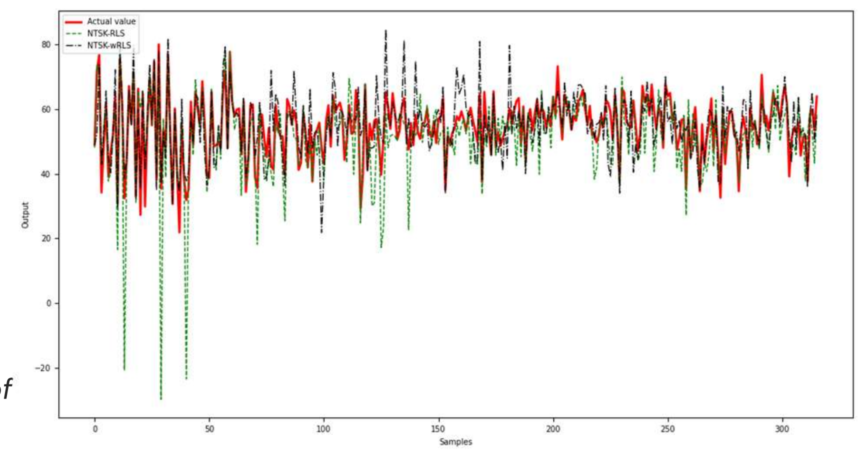


Figure 5.2 – 5.3 – 5.4: *Visualization of Different Model Performances*

COMPARISON & EXPLANATION

Table 5.5: *Summary of Individual Model Performance*

		TSK-MBGD	TSK-GD	NTSK-RLS	NTSK-wRLS
		MODEL			
Parameters	***Rules***	*60*	*30*	*20*	*85*
	Batch Size	*15*	-	-	-
	Lambda	-	-	*2*	-
Validation Metrics	**R^2 Score**	**0.791**	0.718	0.083	0.484
	RMSE	**4.433**	5.123	9.121	6.845
	MAE	**3.401**	3.751	6.802	5.018
	Run Time (s)	908.73	396.31	854.41	561.12
	Run Time (min)	15.15	6.61	14.24	9.35

POTENTIAL REASONS FOR WHICH TSK-MBGD OUTPERFORMS OTHER MODELS

MBGD Optimization: Efficiently handle large data with **smaller batch** updates → Smoother, More **Stable Convergence** = **Reduced variance** of parameter updates & captured **broader data trend** → More controlled, **steady approach** toward the optimal solution = Better overall performance

- **TSK-GD**: Update parameters by processing the **entire** dataset → Potential **overfitting/underfitting** issues
- **NTSK-RLS** & **NTSK-wRLS**: Recursive approach suited for **linear problems**, less effective for complex, nonlinear data

DA-TPA SENSITIVITY ANALYSIS

Purpose Assess feature contribution to predictive accuracy
→ Identify impact VS less relevant features

Process
- Applied ***Min-Max Normalization*** to each column
- Calculated the normalized sum of 3 evaluation metrics
  → **Higher total value = more impactful feature**

Table 5.6: Sensitivity Analysis of R^2 Score, RMSE, and MAE Results across features

Feature	R^2	RMSE	MAE
0	0.7349	4.8628	3.7391
1	0.7312	4.8974	3.7572
2	0.7334	4.8770	3.7493
3	0.7317	4.8926	3.7567
4	0.7315	4.8946	3.7749
5	0.7323	4.8874	3.7594
6	0.7328	4.8823	3.7471
7	0.7299	4.9088	3.7826
8	0.7293	4.9148	3.7844
9	0.7305	4.9032	3.7650
10	0.7315	4.8948	3.7650

Min-Max Normalized →

Feature	R^2	RMSE	MAE	*Sum*
0	1.0000	1.0000	0.9813	***2.9813***
1	0.9008	0.8978	0.9330	***2.7316***
2	0.9593	0.9580	0.9540	***2.8713***
3	0.9146	0.9119	0.9343	***2.7608***
4	0.9089	0.9060	0.8854	***2.7002***
5	0.9295	0.9272	0.9270	***2.7837***
6	0.9443	0.9424	0.9599	***2.8466***
7	0.8679	0.8640	0.8648	***2.5967***
8	0.8504	0.8460	0.8601	***2.5565***
9	0.8841	0.8806	0.9120	***2.6766***
10	0.9081	0.9052	0.9121	***2.7254***

HIGH CONTRIBUTIVE FEATURE

Features with total normalized values > 2.86 ***(9 features)***

Table 5.7: Most Dominant Features

No.	Norm. Val.	Feature No.	Feature
1	2.9813	0	Precipitation
2	2.9131	56	Solar Radiation
3	2.8948	76	Solar Radiation
4	2.8883	107	Snow Water Equivalent
5	2.8801	36	Precipitation
6	2.8785	55	Solar Radiation
7	2.8713	2	Precipitation
8	2.8712	96	Solar Radiation
9	2.8711	103	Solar Radiation

Table 5.8: Composition of Weather Features among Nth Highest Ranked Features

	Nth Highest Ranked Features						
Weather	10	20	30	40	50	60	70
1	**4**	**9**	**11**	**12**	**13**	**16**	**18**
2	**5**	**6**	**11**	**15**	**21**	**27**	**29**
3	1	3	3	3	3	3	6
4	-	**2**	**5**	**10**	**12**	**13**	**16**
5	-	-	-	-	1	1	1
6	-	-	-	-	-	-	-

Highly Contributive Features Identified

- Precipitation
- Solar Radiation
- Maximum Temperature

Key Findings

- Increased impact of **Precipitation** (*Weather 1*) & **Solar Radiation** (*Weather 2*)
- Escalation of **Maximum Temperature** impact observed (*Weather 4*)

LOW CONTRIBUTIVE FEATURE

Features with total normalized values < 1.5 ***(4 features)***

Table 5.9: Least Dominant Features

No.	Norm. Val.	Feature No.	Feature
1	0	117	Snow Water Equivalent
2	0.4882	119	Snow Water Equivalent
3	0.5508	118	Snow Water Equivalent
4	1.3140	139	Snow Water Equivalent
5	2.3321	351	Planting Time
6	2.3487	312	Organic Carbon Density
7	2.3525	249	Vapor Pressure
8	2.3552	286	Bulk Density
9	2.3596	343	Soil Organic Carbon

Table 5.10: Composition of Features among Nth Highest Ranked Features

		Nth Lowest Ranked Features				
		10	20	30	40	50
Soil	*S*	**3**	**6**	**10**	**14**	**19**
Plant	*P*	2	3	5	5	5
Weather	*3*	4	4	4	4	4
	4	-	-	-	-	-
	5	-	**3**	**4**	**8**	**10**
	6	**1**	**4**	**7**	**9**	**12**

Less Impactful Features Identified

- Vapor Pressure
- Minimum Temperature
- Various Soil Features (Sand, Soil Organic Carbon)

Key Findings

- **Minor variance** in evaluation metrics across features
- No single feature drastically outperforms or underperforms
- **Soil** & 2 weather features (***Minimum Temperature*** *(W5)* & ***Vapor Pressure*** *(W6)*) show **minimal** impact

05

CONCLUSIONS

Discussion

Implication

Future Research

DISCUSSION

	Model	
	TSK-MBGD	*CNN-RNN
R^2 Score	79.12%	85.45%-87.09%
RMSE	4.433	4.15-4.91

****CNN-RNN Model** from Khaki, S., Wang, L., & Archontoulis, S.V. (2020). A CNN-RNN Framework for Crop Yield Prediction. Frontiers in Plant Science, 10.*

**Under circumstances: Similar Dataset*

- Performed slightly worse than CNN-RNN
- Potential improvements with parameter tuning and model modifications

IMPLICATION

SCALABILITY & FLEXIBILITY

Effective for non-linear problems

BENEFITS

Enhanced productivity, better resource use, and increased profitability for farmers and policymakers

FUTURE RESEARCH

Enhance Model Accuracy

- Integrate additional data (CO_2, sunlight, pests)
- Refine TSK-MBGD algorithm
- Use cross-validation and advanced ML techniques

Adapt to Changing Conditions

- Incorporate climate change predictions
- Explore time-based models (e.g., RNN, LSTM)

Improve Robustness

- Ensure model generalizability
- Apply risk measurement and Bayesian methods

Reference

[1] Alves, K. S. T. R., De Jesus, C. D., & De Aguiar, E. P. (2024). A new Takagi–Sugeno–Kang model to time series forecasting. Engineering Applications of Artificial Intelligence, 133, 108155.

[2] Khaki, S., Wang, L., & Archontoulis, S.V. (2020). A CNN-RNN Framework for Crop Yield Prediction. Frontiers in Plant Science, 10.

[3] Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks. Frontiers in Plant Science, 10.

[4] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.

[5] Lou, Q., Deng, Z., Xiao, Z., Choi, K., & Wang, S. (2022). Multilabel Takagi-Sugeno-Kang Fuzzy System. IEEE Transactions on Fuzzy Systems, 30(9), 3410–3425.

[6] Everingham, Y., Sexton, J., Skocaj, D. et al. (2016). Accurate Prediction of Sugarcane Yield Using A Random Forest Algorithm. Agron. Sustain. Dev. 36, 27.

[7] Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote sensing of environment, 237, 111599. [8] Di Y, Gao M, Feng F, Li Q, Zhang H. (2022) A New Framework for Winter Wheat Yield Prediction Integrating Deep Learning and Bayesian Optimization. Agronomy. 12(12):3194.

[9] Chen, Chieh-Huang & Lai, Jung-Pin & Chang, Yu-Ming & Lai, Chi-Ju & Pai, Ping-Feng. (2023). A Study of Optimization in Deep Neural Networks for Regression. Electronics. 12. 3071

THANK YOU

Q&A SECTION

DATA DESCRIPTION

Source "A CNN-RNN Framework for Crop Yield Prediction" by Saeed Khaki, Lizhi Wang, and Sotirios Archontoulis (2020)

Loc_ID 1046 locations in 12 states: Indiana, Illinois, Iowa, Minnesota, Missouri, Nebraska, Kansas, North Dakota, South Dakota, Ohio, Kentucky, and Michigan

Year 1980 - 2018

Acronym	Property
bdod	Bulk Density
cec	Cation Exchange Capacity at pH = 7
cfvo	Coarse Fragments
clay	Clay
nitrogen	Total Nitrogen
ocd	Organic Carbon Density
ocs	Organic Carbon Stock
phh2o	pH in H2O
sand	Sand
silt	Silt
soc	Soil Organic Carbon

11 Soil Elements at

6 Diverse Levels of Depth (*0 – 5cm; 5 – 15cm; 15 – 30cm; 30 – 60cm; 60 – 100cm; 100 – 200cm*)

loc_ID	year	yield	W_1_1	W_1_2	W_1_3	W_1_4	W_1_5	W_1_6	W_1_7
0	1980	32.5	0.274725	0	1.615385	0.395604	0.967033	0.736264	1.153846
0	1981	36	0.604396	0	0.043956	0	0.857143	1.824176	0
0	1982	37	2.098901	0.384615	1.681319	0.527473	6.340659	1.593407	1.868132
0	1983	23	0	0	0	1.032967	4.373626	0.351648	0.263736
0	1984	28.5	0	0.043956	0.197802	0.461538	0.142857	0.67033	4.615385
0	1985	39	3.351648	1.56044	1.208791	0	1.956044	2.824176	0.10989
0	1986	36.5	0.131868	0	0	0.142857	2.549451	2.098901	0.857143
0	1987	37	0.098901	2.527473	3.274725	0.065934	0	0.384615	0
0	1988	27	0	0	5.505495	0	2.923077	1.637363	0
0	1989	29	0.978022	0	0.142857	1.252747	1.428571	0.747253	1.758242
0	1990	33.5	1.21978	0	2.021978	1.021978	3.549451	0	5.274725
0	1991	36.5	1.802198	0.615385	0.956044	0.857143	0.054945	0	2.197802
0	1992	38.5	0.582418	0.967033	0	0.648352	0	0.065934	3.945055
0	1993	40	4.593407	2.252747	0.945055	0	0	0.681319	2.417582
0	1994	46	0.582418	0.725275	0.406593	0.230769	0.208791	0.076923	0

P_[1,14]

Planting Time in 14 Planting Date Week

W_[1,6]_[1,52]

[52]: Weeks/Year

[6]: Weather Elements

1. Precipitation
2. Solar Radiation
3. Snow Water Equivalent
4. Maximum Temperature
5. Minimum Temperature
6. Vapor Pressure

11 x 6 = 66 + 14 + 6 x 52 = 312

392 Features