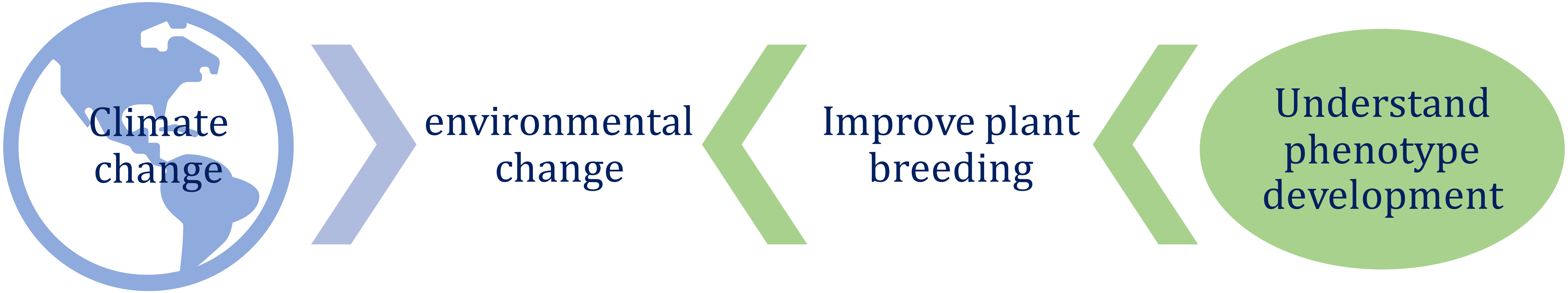
Machine learning for candidate crop growth model classification

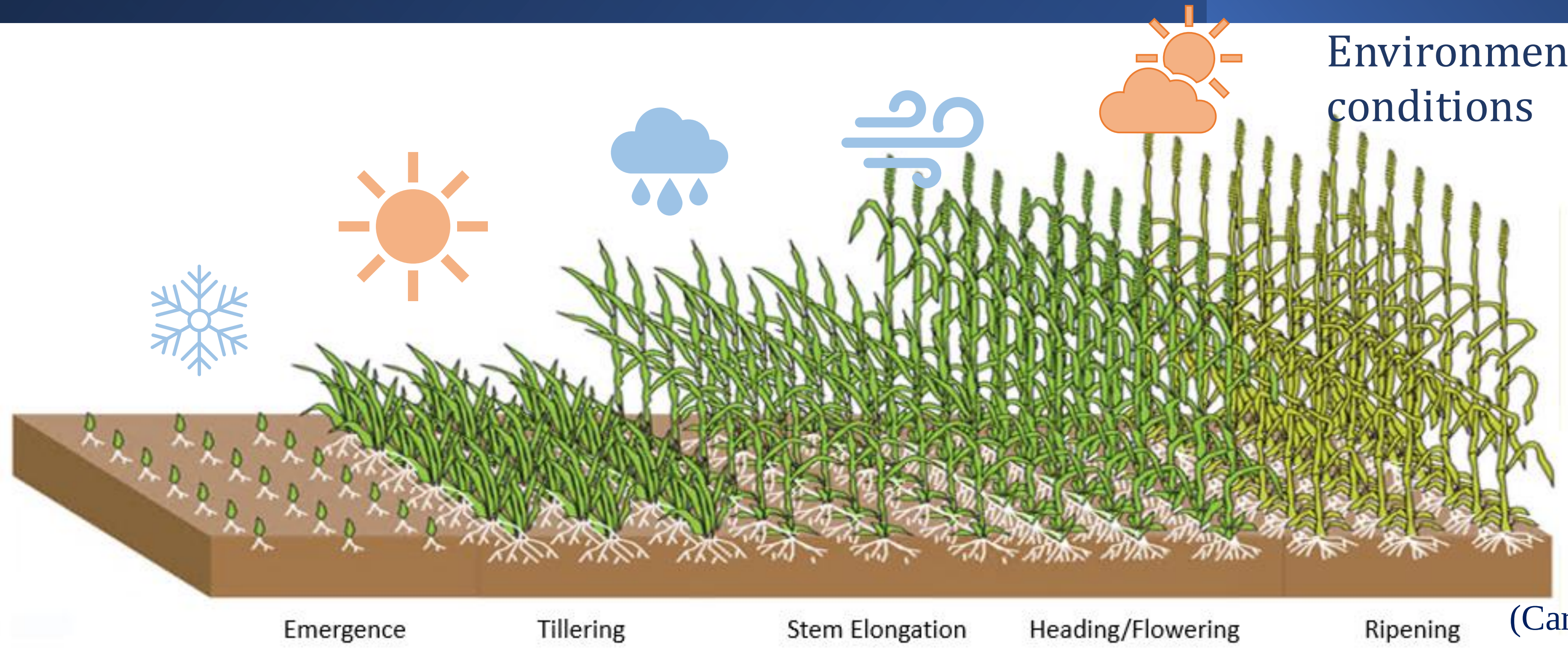
Yingjie Shao

Understanding phenotype development will help plant breeding



Phenotype = f(Genotype, Environment)

Environmental conditions



(Carter 2013)

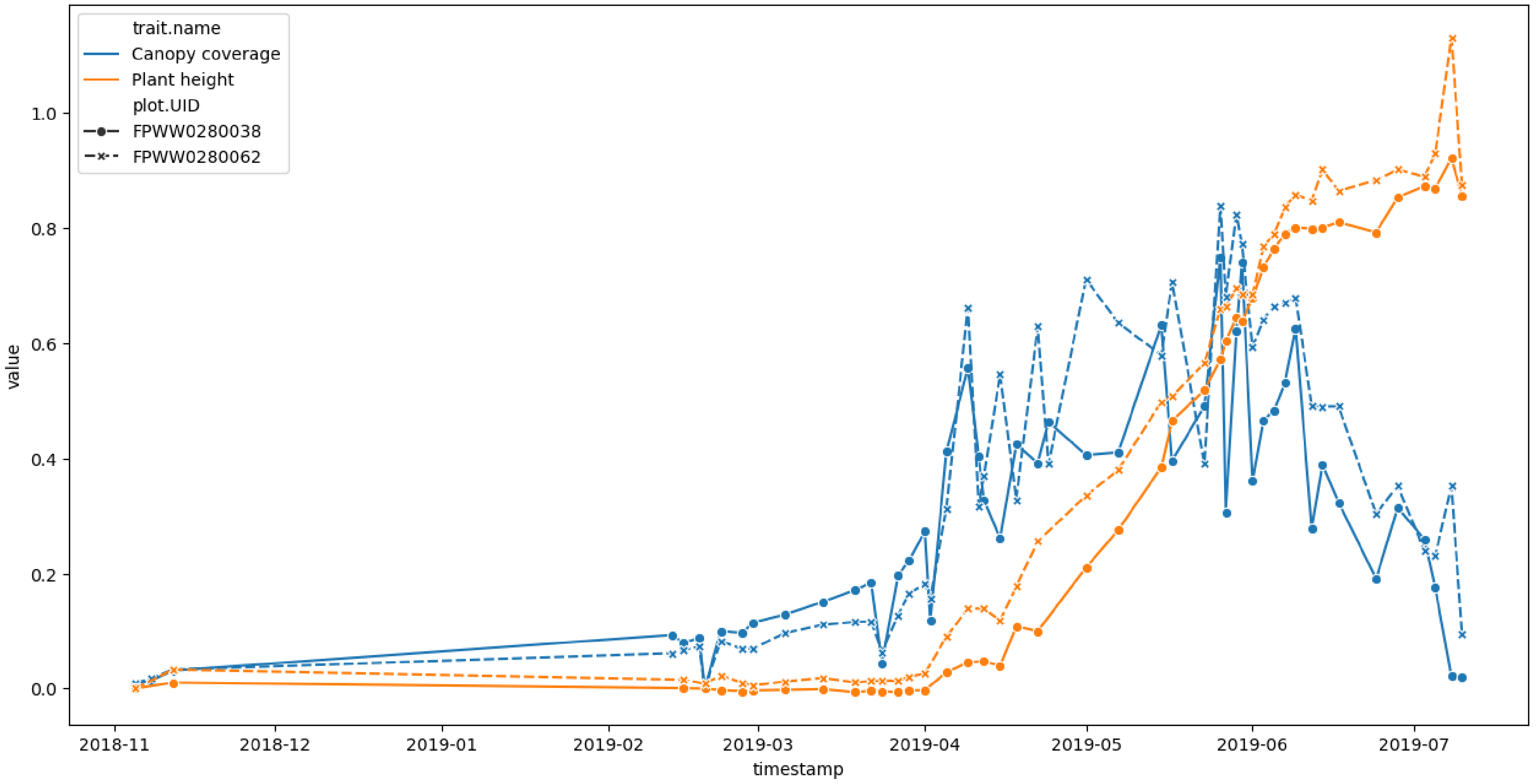
Time

Genotype → Plant breeding

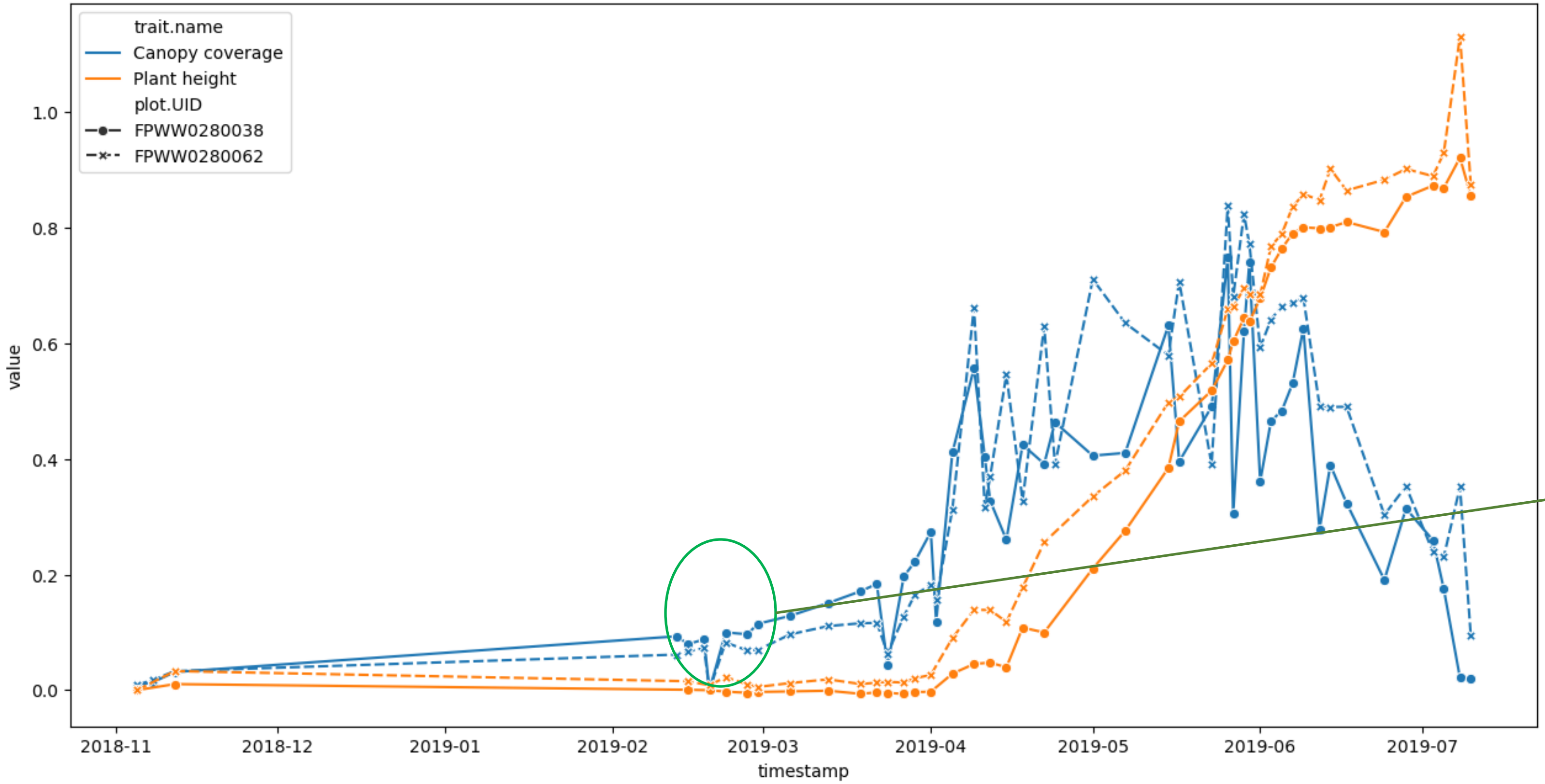
High-throughput phenotyping (HTP) Data

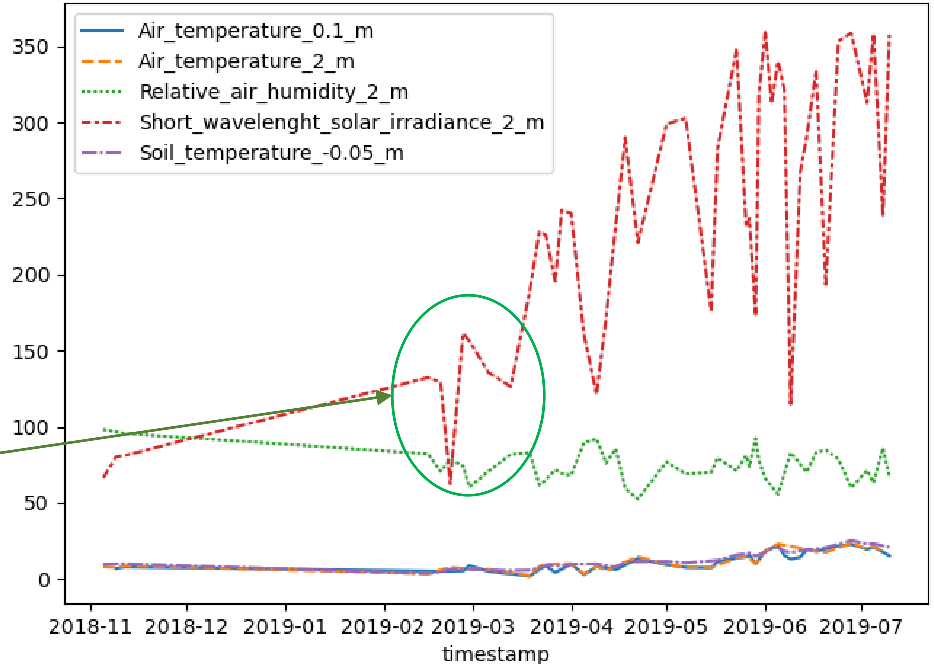
- Non-destructive phenotyping
- Produces multiple traits time series

Time Series Information in Plant Growth



Why Machine learning (ML)?





Extract patterns from time series (crop growth curves), map to non-linear environment effects

Aim

To link phenotype development with the changing environment…

In my MSc thesis, we test how ML works in a simpler case:

Can we use machine learning models to classify the crop growth curve into classes with different environmental limitations?

ODE model for dynamic crop modelling

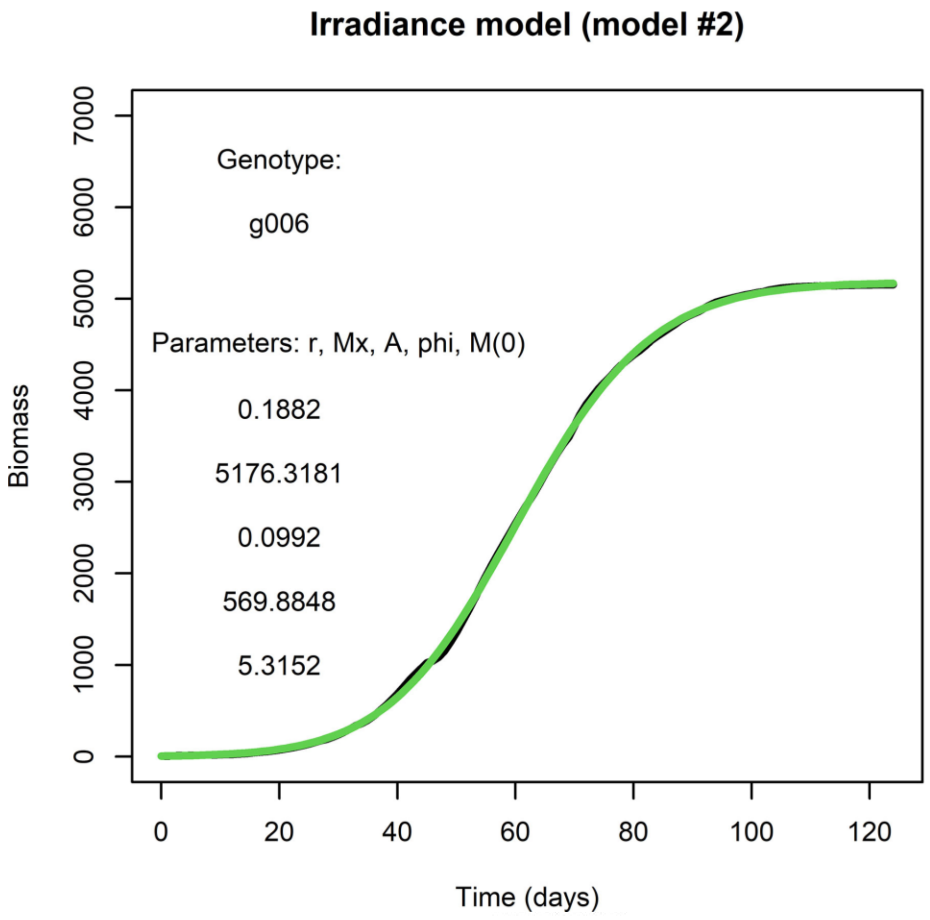
- To generate simulated data…
- Model crop growth as derivatives with respect to time
- Describe dynamic non-linear system

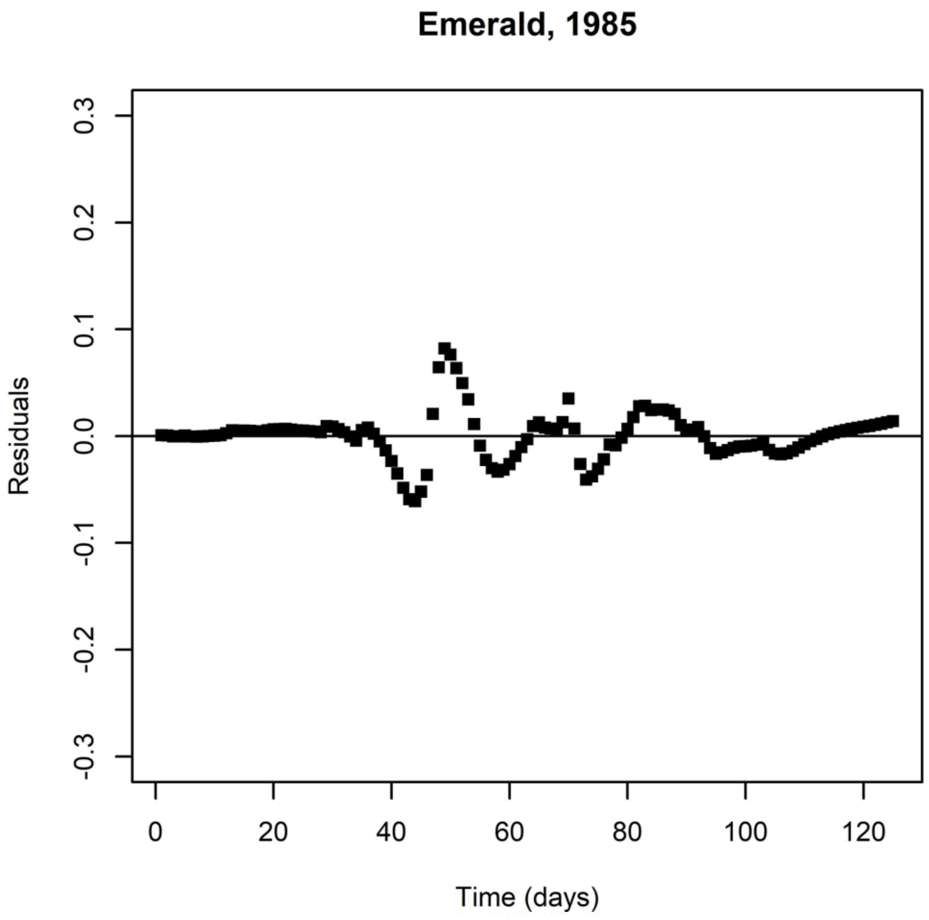
Four types of growth model simulations

Growth Model*	Formula**	Index
Logistic model	$\frac{dM(t)}{dt} = r * M * (1 - \frac{M}{M_{max}})$	1
Irradiance model	$\frac{dM(t)}{dt} = (r + \boxed{(a * sin((2 * \frac{\pi}{365}) * t + \phi))}) * M * (1 - \frac{M}{M_{max}})$	2
Temperature model	$r_{adapt} = (1 + ((exp(\frac{T_{AL}}{temp_t + 273} - \frac{T_{AL}}{T_L} + exp(\frac{T_{AH}}{T_H} - \frac{T_{AH}}{temp_t + 273}))))^{-1}$ $\frac{dM(t)}{dt} = \boxed{r_{adapt}} * r * M * (1 - \frac{M}{M_{max}})$	3
Allee model	$\frac{dM(t)}{dt} = r * M * (1 - \frac{M}{M_{max}}) * (\frac{M}{M + \boxed{Ma}})$	4

(van Voorn, Boer et al. 2023)

ODE models describe plant phenotype development





(van Voorn, Boer et al. 2023)

Fit ODE to daily biomass measurements generated from APSIM-Wheat platform

Data Simulation

- Stochastic differential equation (SDE):
  - Uncertainty of plant growth
  - Stochastic arises from environmental factors and measurement techniques
- Different growth models + different noise types

$$dX_t = \boxed{a(X_t, t)dt} + \boxed{b(X_t, t)}dW_t$$

Four different noise types

Noise type*	Formula
Without noise	$Noise = 0$

Without added noise, it is an Ordinary Differential equation

The highest noise present in the middle of the time series

Time-dependent noise	$Noise = 0.2 * ((2 * \frac{(timelength - t)}{timelength}) * (1 - \frac{(timelength - t)}{timelength})$

The highest noise present when we reach half of the maximum biomass

Biomass-dependent noise	$Noise = 0.2 * ((2 * (M_{max} - x)/M_{max}) * (1 - (M_{max} - x)/M_{max}))$
Independent noise	$Noise \sim N(0, 0.25)^{**}$

Input Data set

Input X: 4*4 SDE models, 300 samples per SDE model

For LSTM model

Daily biomass shape = (120,1200,1)

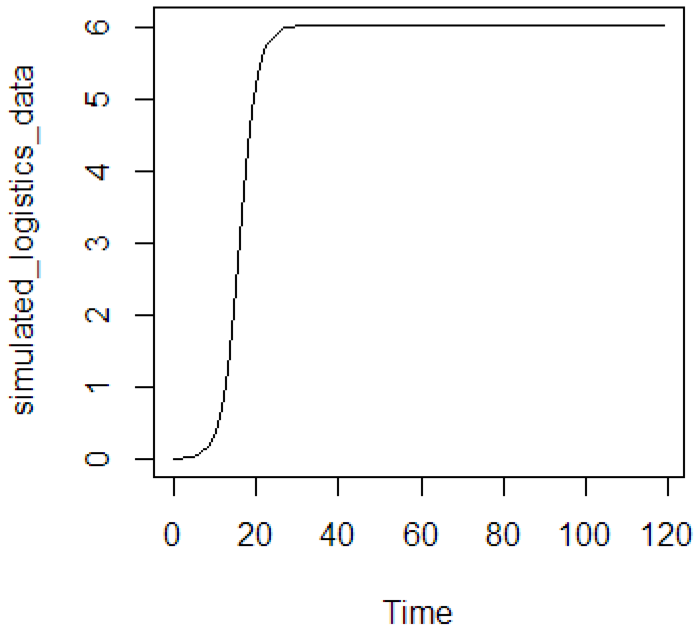
Daily biomass + the derivative of biomass shape = (120,1200,2)

For CNN model

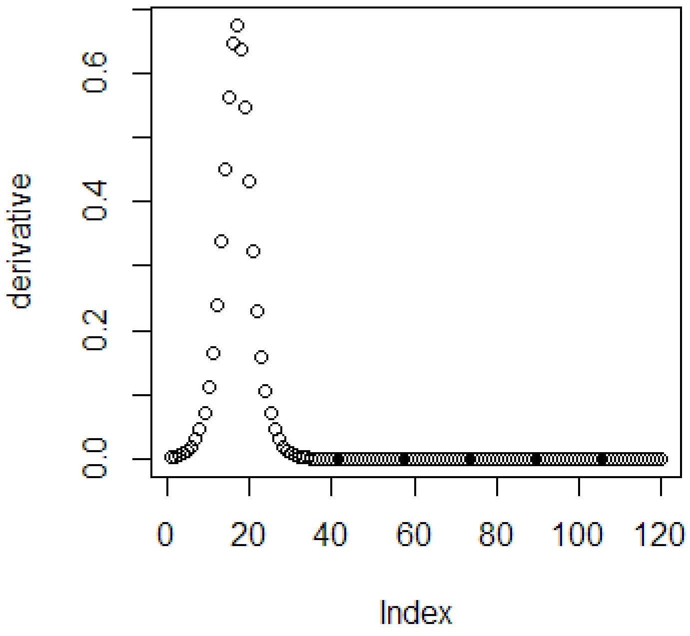
The image of Daily biomass + the derivative of biomass shape = (1200,2,256,256)

Input Y: growth model types （4 classes labels）

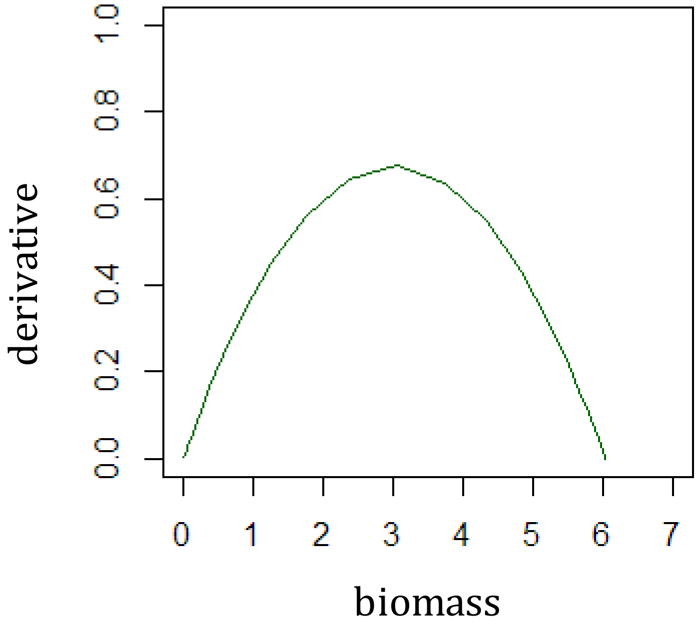
Example of curves without noise



Biomass Growth curve

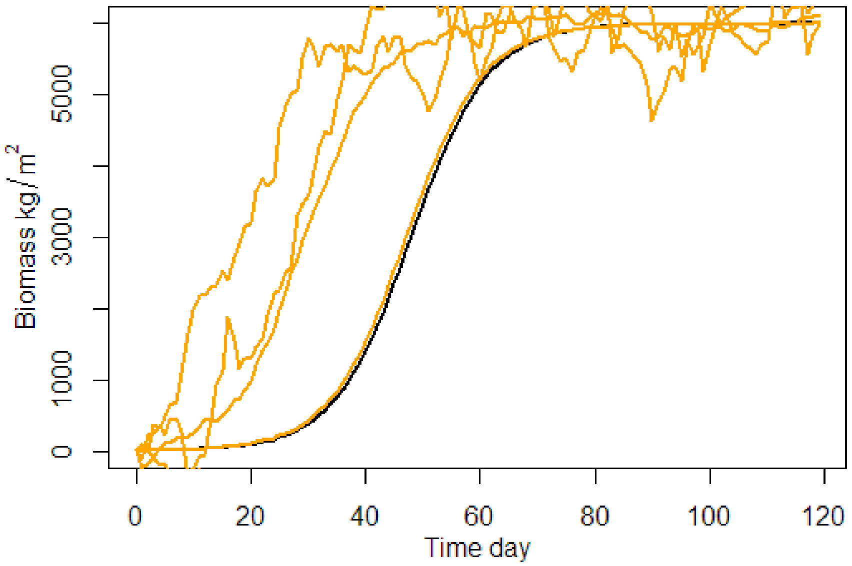


Biomass derivative against days

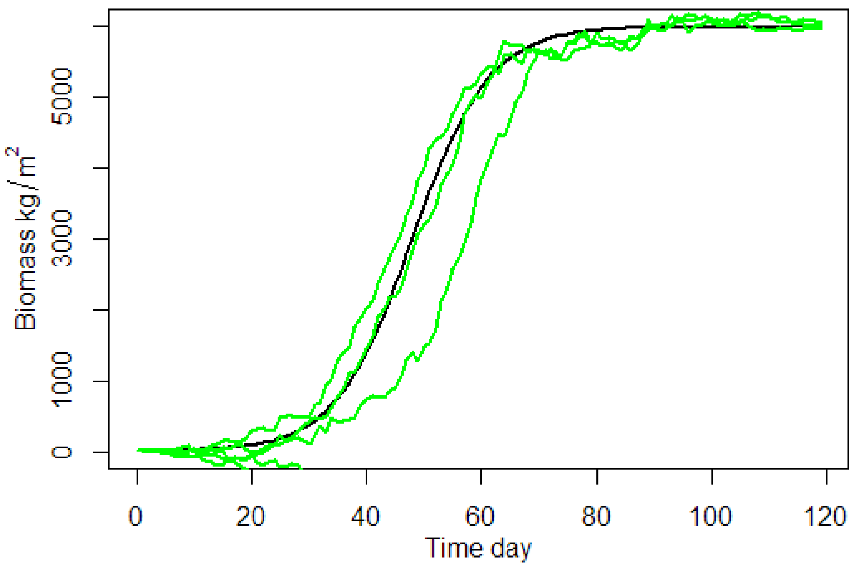


Biomass derivative against biomass

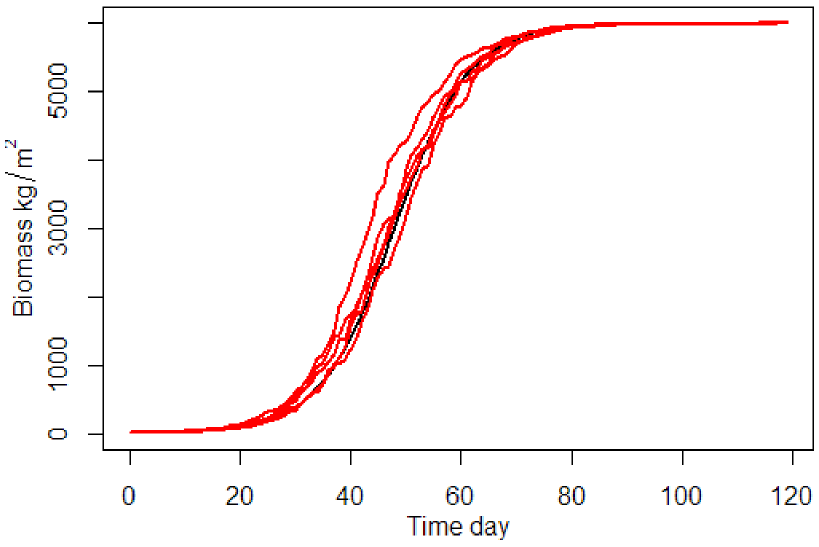
Biomass growth curve for different noise types



Independent noise

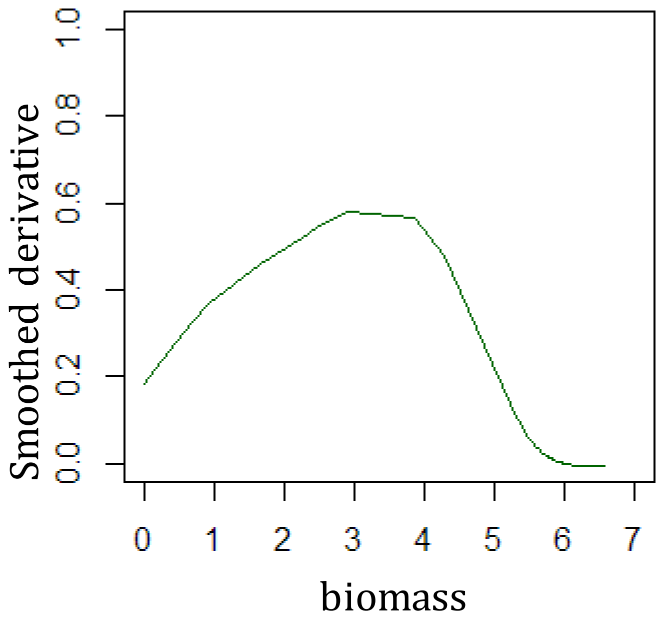


Time-dependent noise

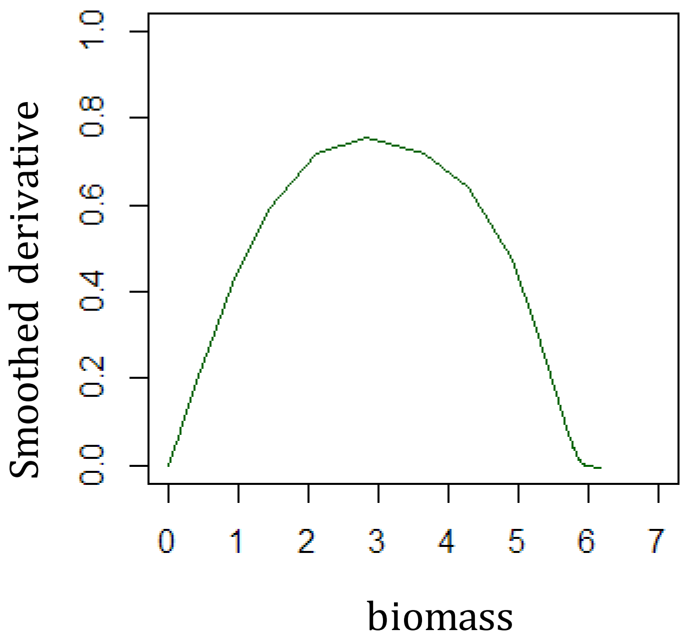


Biomass-dependent noise

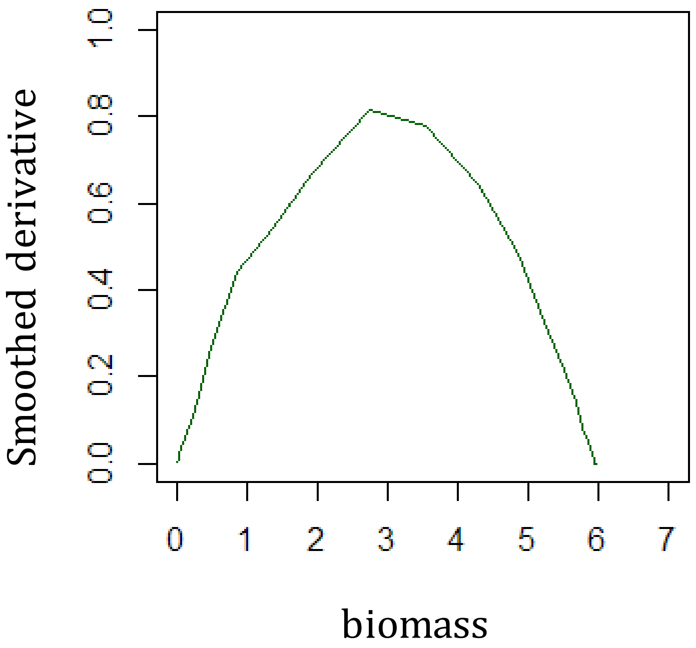
Biomass derivative against biomass for different noise types



Independent noise



Time-dependent noise



Biomass-dependent noise

Smoothed derivative based on pspline (nkots=36)

Input Data preprocessing

Input X:

split based on noise type -> models are trained separately for four noise types.

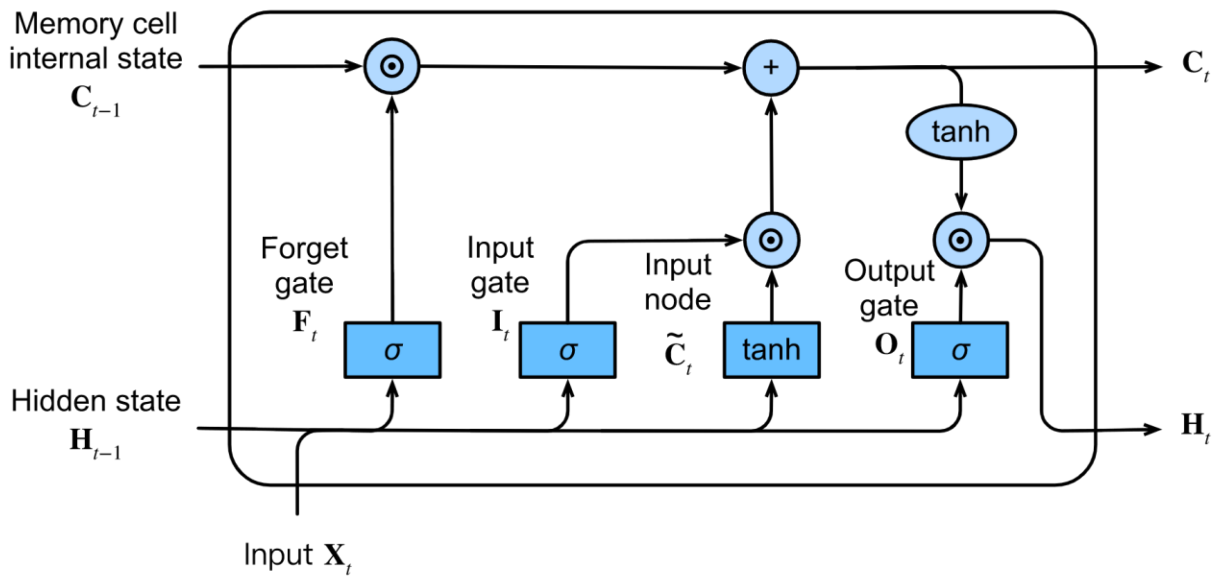
Scaling before feeding into neural networks $z = \frac{x-\mu}{s}$

Input Y: One-hot encoded growth model types

[1,0,0,0]
[0,1,0,0]
[0,0,1,0]
[0,0,0,1]

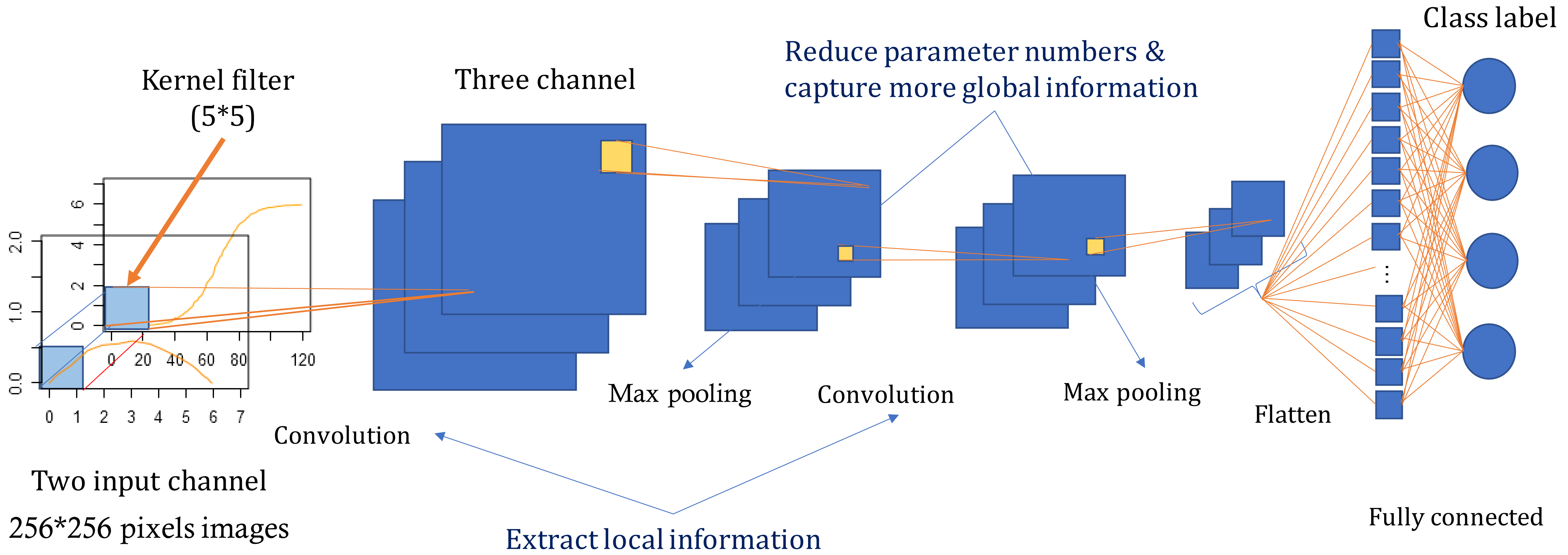
LSTM Model

- Two LSTM layer + fully connected layer
- Time series sequences as input

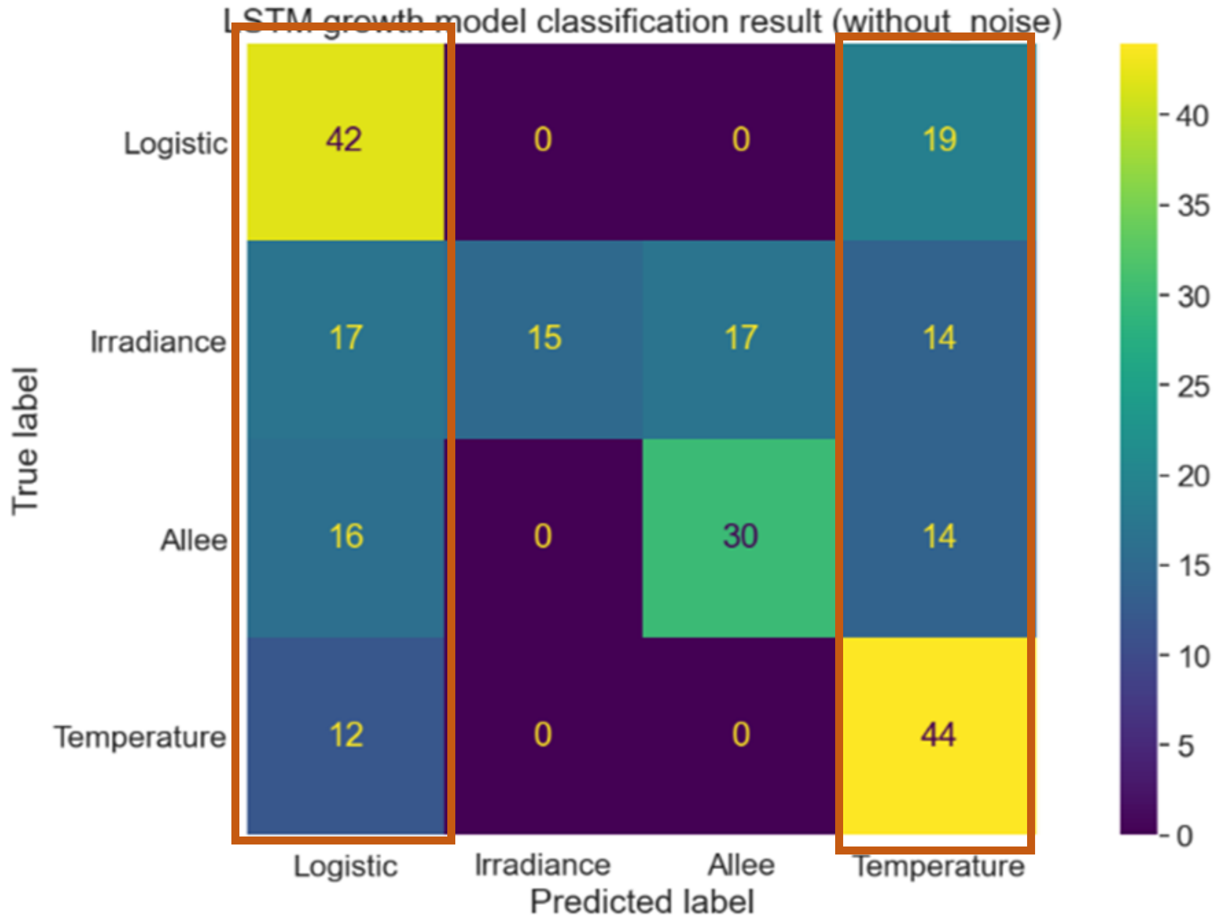


(Zhang, Lipton et al. 2021)

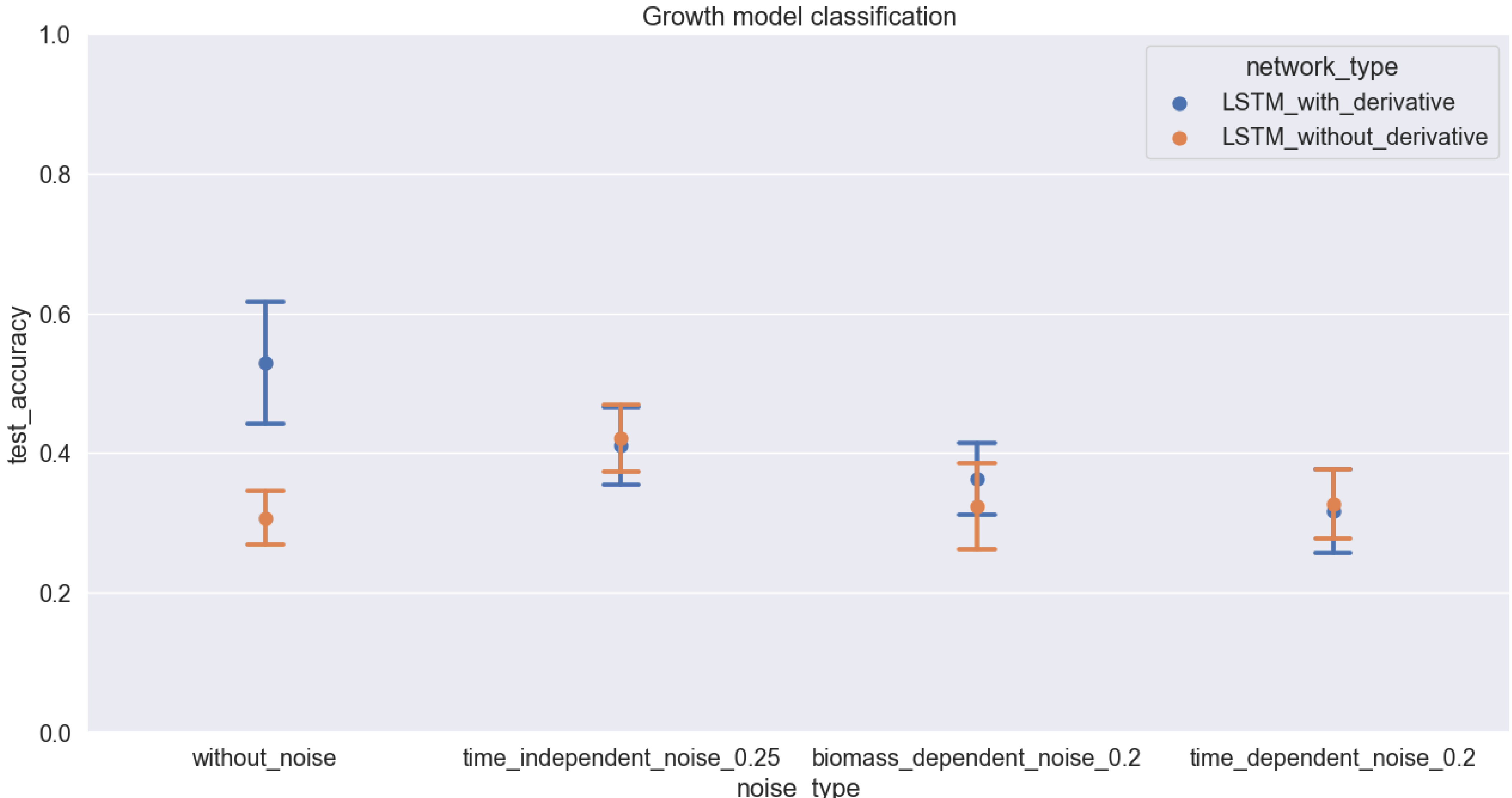
CNN Model


Kernel filter
(5*5)
Three channel
Reduce parameter numbers &
capture more global information
Class label
0 20 40 60 80 120
0 1 2 3 4 5 6 7
0.0 1.0 2.0
0 2 4 6
Max pooling
Convolution
Max pooling
Flatten
Convolution
Two input channel
256*256 pixels images
Extract local information
Fully connected

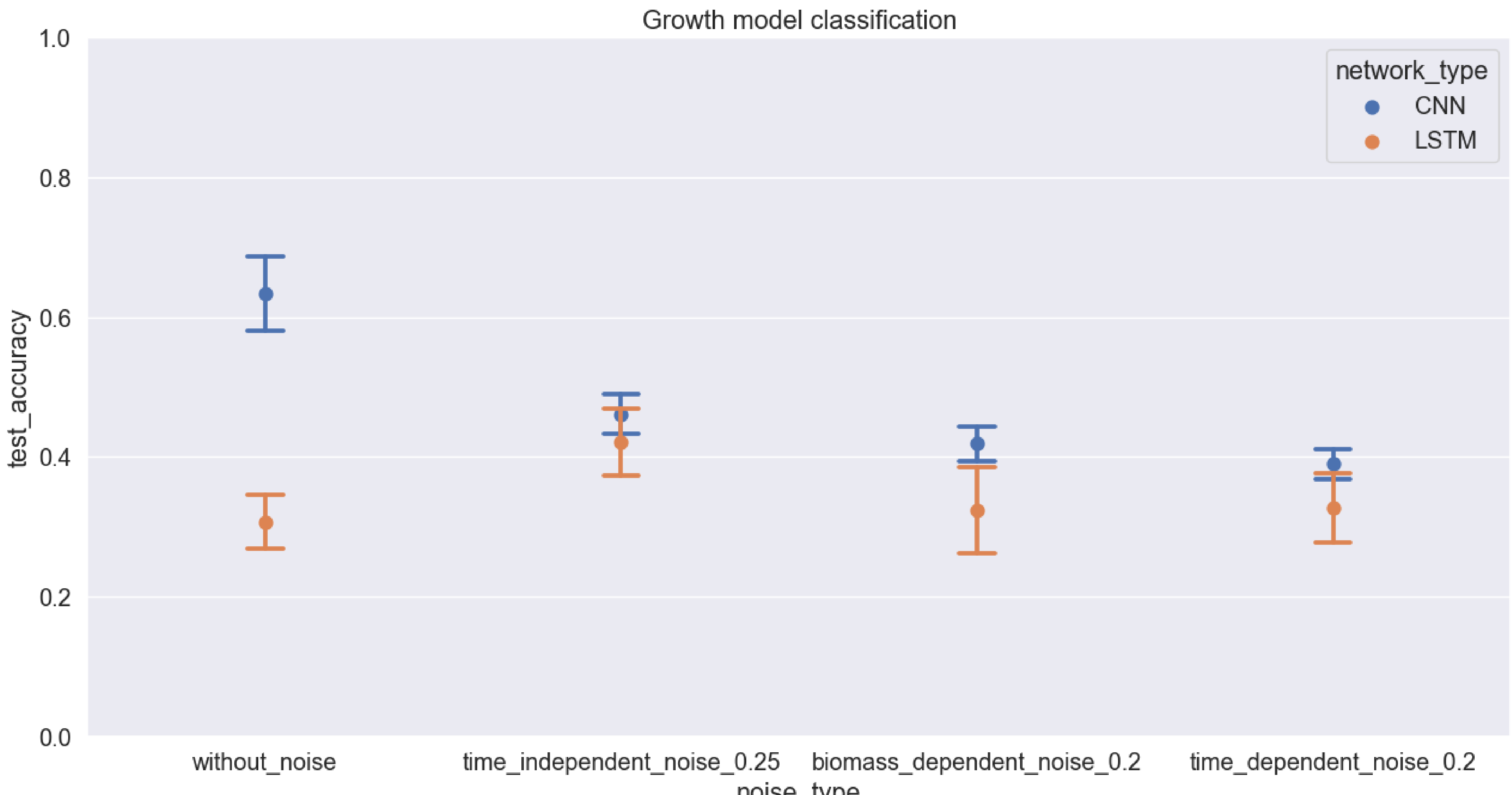
The model tends to classify samples into logistic or temperature class



Added Derivate Information Improves Accuracy



Growth Model Classification Comparison



Conclusion& Discussion

Adding derivative information improves prediction accuracy ->

Extracted features from differential equations and combined them with raw data

Both models give relatively high accuracy ->

Combine CNN with LSTM and use previous knowledge to guide hyperparameter selection

Future Plan

Physical knowledge + neural network

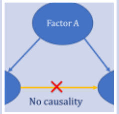
↓

CausalML

Knowledge gaps

Quantify dynamic genome effects within different environments

Combining process-based and data-driven models in plant studies



Causal discovery and estimation for time series (HTP data)

Objectives

- Build accurate plant growth forecasting models by combining process-based models with neural networks
- Quantify G*E interactions in a dynamic system of plant growth
- Discover causal relationships between phenotypes and environmental factors

Thank you!

Questions?