Genomic Prediction of Traits Using Convolutional Neural Networks

Emily Bolger[1], Christina Azodi[2], Shin-Han Shiu[2]
[1]Moravian College, [2]Michigan State University, Department of Plant Biology
iCER ACRES REU, Michigan State University

Convolution Neural Networks do not perform as well as regression-based or classical machine learning models at predicting trait values from genetic information.

OBJECTIVES

Using Convolutional Neural Networks (CNNs) to predict trait values in Spruce, specifically Height (HT), Wood Density (DE), and Diameter at Breast Height (DBH), and comparing the accuracy to a variety of other predictive models.

BACKGROUND

What is Genomic Prediction (GP)?

- Using genetic information to predict trait values

How do we do it?

- By training predictive models, like Deep Learning

Why do we use it?

- Save plant and animal breeders time and money
  - Identifying important genetic markers before the breeding process allows for focus on particular combinations that will be most fruitful, resilient, diverse
- By interpreting our deep learning models, we can better understand how traits are controlled by DNA

DATA

What genetic information do we use?

- Most of the sequences in the genome are identical between different varieties of the same species
  - Small snippets where these sequences differ are called *genetic markers*
  - Some of these genetic markers will be associated with variations in the trait values of a plant

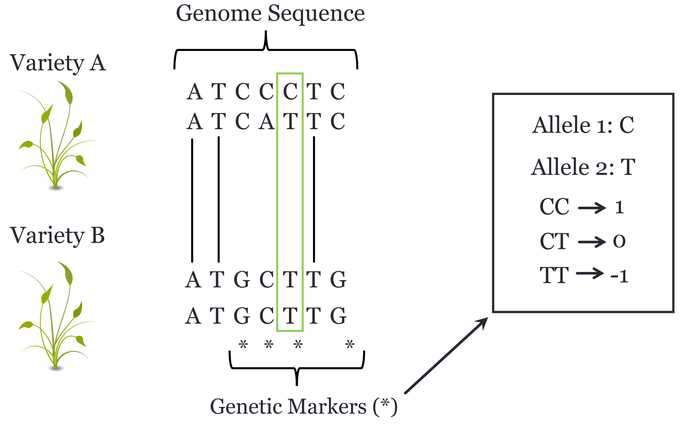

Figure 1: Example of how to recognize genetic markers within DNA sequences and how we label the allele combinations with 1,0,-1s. Our dataset has 6930 genetic markers and 1205 spruce varieties. This ratio of a large number of markers to a small number of individuals often creates a challenge in Genomic Prediction.

METHODS

Our Convolutional Neural Network

- Deep Learning tool typically used for image classification
- As the model trains, various layers:
  - Assign weights based on feature location, decrease input size, avoid overfitting
- Model learns how the pattern of 1,0,-1s is associated to a specific trait value
- Input: Genetic Data (1,0,-1s), Output: Single predicted trait values

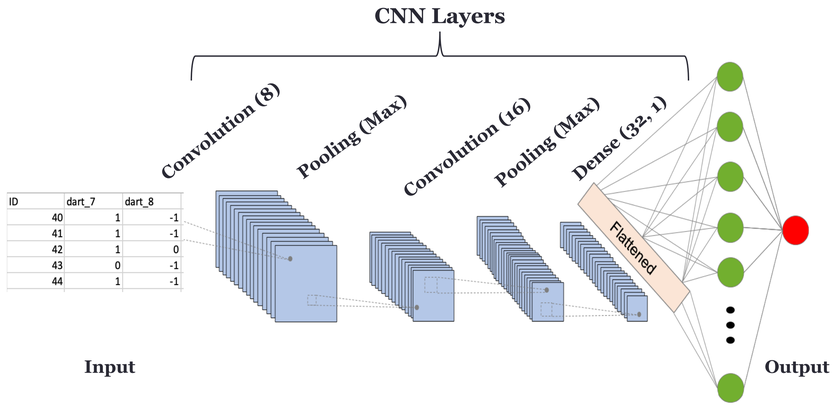

Figure 2: Visual based on the structure (8-16-32-1) of our Convolutional Neural Network, built in TensorFlow. Structure based off of the network proposed in DeepGS model(Saha, Sumit, 2018; Ma, Wenlong, et al., 2017, Vol 1. Pg 3).

PARAMETER SELECTION

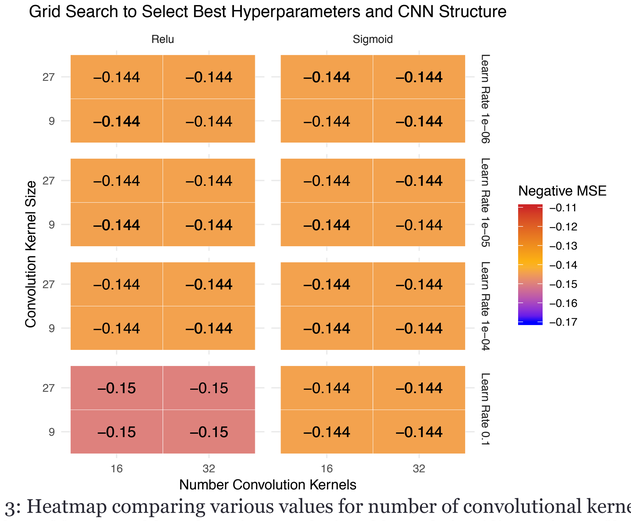

Figure 3: Heatmap comparing various values for number of convolutional kernels (~number of features able to learn), convolutional kernel size (dimensions of kernel), learning rate (how quickly model learns), and activation layer type (apply nonlinearity) used in Randomized Grid Search. Accuracy measured with Negative Mean Square Error. Appears that model is insensitive to parameter changes.

MODEL PERFORMANCE

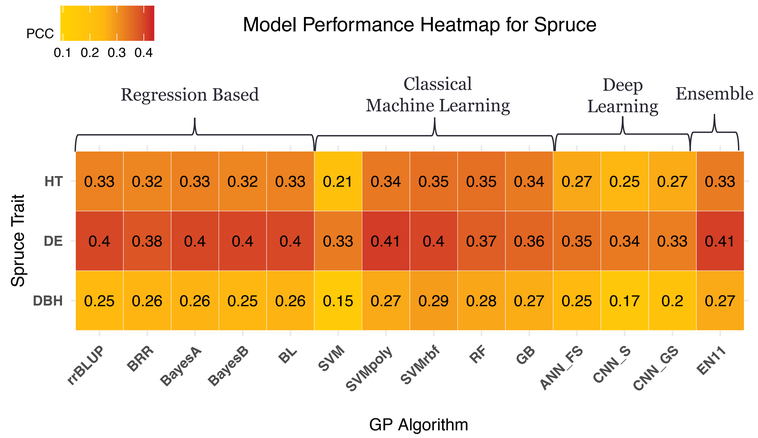

Figure 4: Heatmap (Azodi, Christina B., et al., 2019, Vol.1, Pg 6) depicting the accuracy of 14 models in predicting HT, DE, and DBH in Spruce. Pearson's Correlation Coefficient (PCC) is used as measure of accuracy. Models that more accurately predict trait values are indicated with a red background.

Here, CNN_S is our simple model with structure 8-1 and CNN_GS is our model based on DeepGS framework with structure 8-16-32-1.

Models from left to right: rrBLUP, ridge regression Best Linear Unbiased Predictor; BRR, Bayesian Ridge Regression; BayesA; BayesB; BL, Bayesian LASSO; SVR, Support Vector Machine (kernel type: poly, polynomial; rbf, radial basis function); RF, Random Forest; GB, Gradient Tree Boosting; ANN_FS, Artificial Neural Network with Feature Selection; CNN_S, Convolutional Neural Network Simple Frame Work; CNN_GS, Convolutional Neural Network based on DeepGS Model; EN11; Ensemble Model with all 11 models (all except CNN).

CONCLUSIONS

Where does the CNN stand?

- CNN less accurate than Regression Based and Classical Machine Learning models for all three traits
  - Performs slightly worse than the other deep learning model, a fully connected Artificial Neural Network with Feature Selection

Next Steps

- Introduce Feature Selection (focus on genetic markers that most heavily contribute to trait of study) to see if that addresses the marker to number of individuals ratio and improves performance
- Order the genetic markers in input data to see if this helps CNN with pattern recognition
- Apply CNN model to 5 other species across 18 different traits and compare the accuracy to 11 other major predictive models

ACKNOWLEDGEMENTS

Thank you to all in the Shiu Lab, especially Christina Azodi, for their guidance and patience this summer. Thank you to Michigan State University's CMSE Department and the iCER ACRES REU program for giving me this opportunity.

NSF ACRES REU – OAC1560168