The region-based soybean yield prediction using stochastic, machine learning, and deep learning methods

Authors: Alice Li, Shawn Lang, Ziang Xu

December 29, 2022

Abstract

This project looks into what is the most feasible strategy for soybean yield prediction. We used data combined from the USDA National Statistics Service, Economic Research Service, and NOAA climate data. In our estimation part, we compared three types of estimation models: classic statistical linear regression model (LR), machine learning random forest estimator (RF), and deep learning feed-forward neural network estimator (FFNN). We build the integral model with the whole dataset as well as the region-based model with sub-datasets separated by the USDA agricultural regions with each of the three estimators. We made model evaluations with a cross-validation method on all of our models. For each model, we calculated the model performance on the hold-out test dataset error by getting the mean average error (MAE), root-mean-square deviation (RMSE), and proportion of the variance explained by the model (R^2). We compared region-based estimation models vs. the integral estimation model to see whether the region-based estimation models can facilitate performance. We also compared all the region-based estimation models to see the model performance characteristics for each region. As a result, we noticed that region-based estimators are more accurate than the integral estimator. Among all the models, random forest regression performs the best with the region-based dataset; FFNN performs better on the prediction of the whole dataset but performs worst on the region-based dataset.

1 Introduction

Crop yield is a crucial indicator in agriculture and is closely related to livelihood and the global economy. Soybean, as one of the most dominant crops in the United States, is widely used in food and industrial applications, including making fermented and unfermented food and extracting soy vegetable oil. It is one of the richest and cheapest sources of protein and is a staple part of the daily diets of people and animals in most parts of the world [1]. The United States serves as the world's leading soybean producer, which supplies 34% of global annual soybean production, and is the second-leading exporter of soybean. Soybeans comprise about 90 percent of U.S. oil seed production. By 2020, the U.S. soybean planted acreage achieved 90 million acres, 4,135 million bushels production, and $46.1 billion in cash flow [2].

Soybean has a large impact on the U.S. economy. The Economic Impact of U.S. Soybeans & End Products on the U.S. economy shows that the total economic impact on the U.S. economy from the soybean sector averaged $115.8 billion per year including $7.96 billion from crushing – the equivalent of more than 0.65 percent of the U.S. gross domestic product (GDP), and up to nine percent of the GDP for certain states. In recent years, the majority of soybean crops are genetically modified for resistance to the herbicide glyphosate in the United States [2].

Climate conditions can have large impacts on soybean production, like most agricultural plantations. The best soybean yields occur on well-drained, but not sandy, soils having a pH of 6.5 or above. The critical stage for soybean yield is in August and soils that typically dry out in August will have disappointing yields. Most U.S. soybean-producing regions are rain-fed, making them highly vulnerable to extreme weather conditions. Weather condition during the growing season is critical for soybean yields. Variability in growing season precipitation and temperature induced by climate change are important constraints in crop production [3].

Moreover, various aspects of soybean production generate greenhouse gases that contribute to climate change. The growing acres of soybean agriculture largely changed the soil composition and caused deforesta-

tion, soil erosion, decreasing biodiversity, increasing carbon emission, and straining water resources, which all worsen the global climate change [3].

Based on these reasons, a feasible and accurate yield prediction strategy is necessary for yield mapping, soybean market pricing and planning, harvest management, and estimating its impact on the soil as well as the climate. In this research, we aim to find out how to approach an appropriate method for the estimation of soybean yield by understanding more deeply about soybean yield and its relationship with possible impacts, including weather, chemical fertilizer application, the percentage of genetically modified, etc. USDA has made clear divisions of agricultural regions of the entire soybean cultivated regions based on their geological characteristics, called USDA agricultural region code [2]. It is novel and intriguing to look at the effectiveness of statistical, machine learning, and deep learning methods in predicting soybean yield based on these region codes. We are also looking into whether soybean plantations in different agricultural regions have different performances and how that division would affect our prediction of soybean yield.

2 Research Question

In this research, we aimed to answer: What is the most feasible strategy for soybean yield prediction?

3 Data and Methods

The dataset was collected by combining data from the USDA National Statistics Service, Economic Research Service, and NOAA climate data. It is comprised of features in a county's suitability for growing soybean, genetic modification (GM) uptake, irrigation, and other intensive inputs over the period 1990-2017 [2, 4].

The raw dataset describes 1751 different counties' soybean cultivation and production information in 28 years (1990 to 2017). Therein, it contains 49028 instances and 13 variables, in which each instance contains the cultivation and production information of a county in one specific year. This project aimed to predict the soybean yield and thirteen features were considered essential to the soybean yield, which can be divided into three categories. The first category contains categorical features identifying the general cultivation information, including year, county code, and USDA agricultural region code. USDA agricultural region code is the standard region code generated by USDA in the criteria of different geographical characteristics and divides all U.S. soybean cultivated lands into six distinct sections (region codes defined as R1 to R6). The second category contains soybean cultivating details, including fertilizers application rate (potash fertilizer application rate, nitrogen fertilizer application rate, phosphate fertilizer application rate, and genetic modification application rate), irrigated percentage, chemicals expenditure, and machinery expenditure. The last category focuses on the climate data for each county, which includes the precipitation and temperature in the growing season.

The correlation of all ten numerical features in categories 2 and 3 was plotted in Fig. 1A [5]. All these ten features are not highly correlated. Besides, as shown in Fig. 1B, soybean yields per acre in different USDA agricultural region codes are various, representing the importance of finding the best model for predicting soybean yield in different regions. Following the schematic workflow shown in Fig. 2, in the preprocessing step, we removed the duplicated instances, instances with missing values, and all the outliers more than three standard deviations threshold in this dataset [5]. As analyzed above, the ten most vital features and the target, soybean yield, were kept. We performed a 5-fold cross-validation with python based package on the dataset to train selected models [6, 7].

On the ground of the research question, two estimation tasks were built on the dataset. The first task aimed to estimate the yield using the integral dataset (all U.S. soybean cultivated regions). In the second task, we divided the dataset into six smaller datasets based on the USDA agricultural region code (R1 to R6). We estimated the soybean yield in each small dataset. The estimation task was justified by comparing different models among statistical, machine learning, and deep-learning models, where we selected multiple linear regression (LR), random forest regression (RF), and feed-forward neural network (FFNN), respectively. Regarding the FFNN model, we set a total of three types of layers, which are the input layer (input shape equals 32), the hidden layer, and the output layer. The model initializer and the activation function were set to be normal and rectified linear units(ReLU), respectively. We set nine for our input dimension because there are nine features included in our model. Besides, we have three hidden layers in our model, where

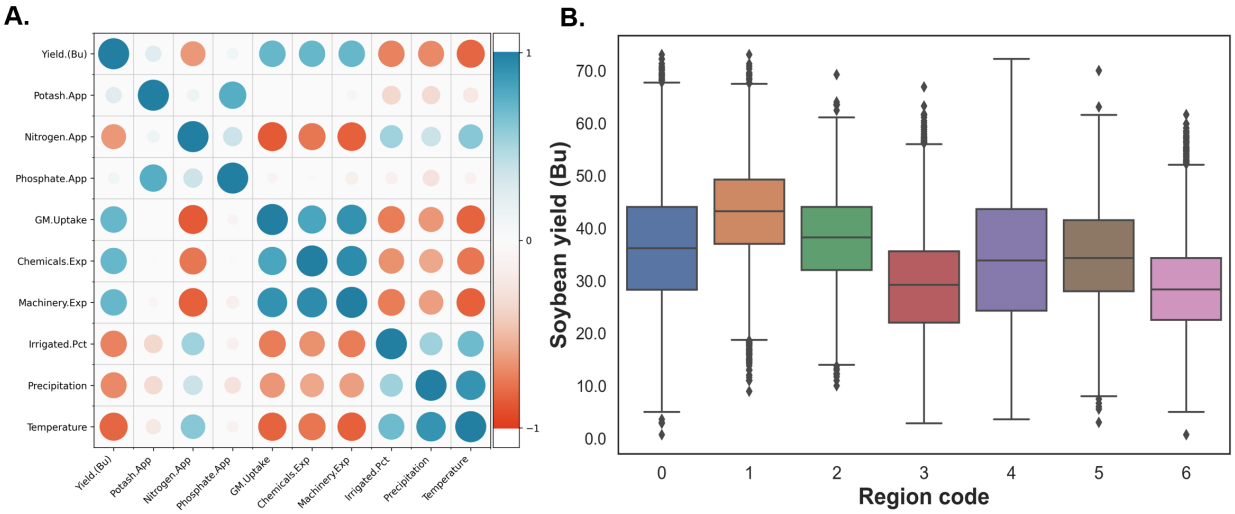

Figure 1: Visualization of the soybean data set. (A) The correlation graph between 10 hub features and the target, soybean yield. (B) The boxplot of soybean yield per acre with different agricultural region codes, where region code 0 here accounts for the integral dataset (all U.S. soybean cultivated region) and the rest account for the USDA agricultural region codes from 1 to 6.

each had an input shape of 64. The output layer has shape one since only one estimation result was aimed here [8]. For the random forest model, it had a max depth of 5, where other parameters followed the default settings. We chose the max depth to be five because a depth of five can balance the accuracy and over-fitting, and require moderate computational time.

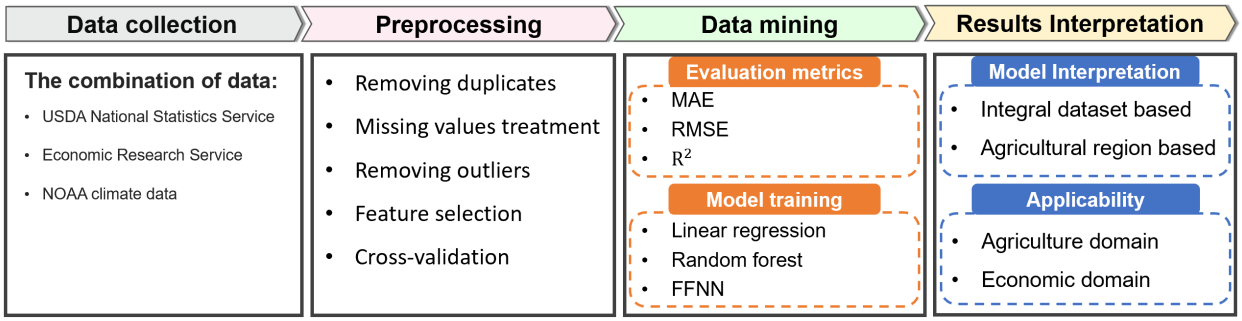

Figure 2: Schematic illustration of the workflow of the data mining task about soybean yield prediction. Data were first collected from three different datasets and preprocessed. We trained models with different underlying methodologies based on these data and evaluated them by three different evaluation metrics. The results obtained were interpreted and future insights were proposed.

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{1}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{2}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3}$$

Moreover, to measure each model's performance, we used three evaluation matrices, including mean absolute error (MAE, Eqn. 1), root mean square error (RMSE, Eqn. 2), and R-squared error (R^2, Eqn. 3). All the models and evaluation metrics mentioned above are python-based [6, 8]. The naive MAE and naive RMSE of each test dataset was also computed to gain more insights into the effectiveness of these models.

4 Results

The correlation plot among the ten selected features and the soybean yield per acre provides us with some insights. The correlation plot (Fig. 1A) displays a clear negative correlation between soybean yield and the temperature increase in the soybean growing season and a negative correlation between soybean yield and the precipitation increase in the soybean growing season. There are positive correlations between soybean yield and generic modification technique uptake, expenditures in chemicals, and expenditures in machinery. This means that investment in chemicals, machinery, and gm update are related to soybean yield. Some out-of-common-sense patterns shown in our correlation plot include a negative relationship between nitrogen application rate and the soybean yield and a negative relationship between irrigation percent and the soybean yield. As these relationships are relatively weak, we still need more data to understand and rationale them.

MAE Evaluation				
Region code	**Naïve**	**LR**	**RF**	**FFNN**
R0	9.019 ± 0.078	7.366 ± 0.053	6.271 ± 0.018	5.548 ± 0.077
R1	7.133 ± 0.087	5.709 ± 0.076	5.036 ± 0.026	4.996 ± 0.109
R2	6.895 ± 0.112	6.261 ± 0.159	5.297 ± 0.169	5.562 ± 0.198
R3	8.200 ± 0.218	6.396 ± 0.109	5.310 ± 0.146	10.617 ± 10.196
R4	10.592 ± 0.209	7.728 ± 0.366	6.683 ± 0.257	7.294 ± 0.818
R5	7.863 ± 0.155	6.210 ± 0.138	5.317 ± 0.142	5.418 ± 0.199
R6	6.851 ± 0.101	6.107 ± 0.039	5.437 ± 0.047	5.724 ± 0.155
RMSE Evaluation				
Region code	**Naïve**	**LR**	**RF**	**FFNN**
R0	11.055 ± 0.066	9.247 ± 0.075	7.993 ± 0.024	7.266 ± 0.101
R1	8.897 ± 0.106	7.307 ± 0.088	6.397 ± 0.074	6.758 ± 0.136
R2	8.588 ± 0.163	7.927 ± 0.231	6.763 ± 0.234	7.097 ± 0.263
R3	10.408 ± 0.197	8.419 ± 0.174	7.007 ± 0.142	12.905 ± 9.931
R4	12.857 ± 0.248	9.624 ± 0.384	8.408 ± 0.358	9.424 ± 1.087
R5	9.722 ± 0.160	7.869 ± 0.113	6.860 ± 0.131	7.078 ± 0.167
R6	8.559 ± 0.145	7.693 ± 0.093	6.888 ± 0.068	7.600 ± 0.309
R^2 Evaluation				
Region code	**Naïve**	**LR**	**RF**	**FFNN**
R0	/	0.300 ± 0.009	0.477 ± 0.006	0.567 ± 0.013
R1	/	0.325 ± 0.014	0.483 ± 0.014	0.422 ± 0.031
R2	/	0.148 ± 0.022	0.380 ± 0.029	0.317 ± 0.036
R3	/	0.345 ± 0.023	0.547 ± 0.017	-1.278 ± 3.615
R4	/	0.439 ± 0.046	0.572 ± 0.034	0.459 ± 0.137
R5	/	0.345 ± 0.017	0.502 ± 0.012	0.470 ± 0.012
R6	/	0.19 ± 0.013	0.352 ± 0.017	0.211 ± 0.058

Table 1: Performance evaluation of statistical, machine learning, and deep-learning models with three evaluation metrics, including MAE, RMSE, and R^2.

Furthermore, we look into the six agricultural regions and their distinguished characteristics related to soybean yield. we dive into the plantation condition differences between each agricultural region as well as the characteristics of each region's soybean production. We noticed a clear pattern in the difference between the agricultural regions and their temperature and precipitation conditions for soybean plantations during the growing season. For instance, Region 1 has the most suitable growing season temperature and

precipitation conditions for soybean yield and has comparatively higher soybean production. Region 3 in general has the lowest temperature for the soybean during the growing season and this region has some really large soybean plantations in certain counties. Region 6 has in general higher amount of precipitation than other regions.

In order to gain some insights into how the soybean plantation characteristics have evolved over time, we also took look into conditions with respect to years. Comparing changes between 1990 to 2017, region 6 has had an increase in large-scale plantations shown by the changes in soybean planted acres for each county as well as the number of the county that was involved in the soybean plantation activity. Another major change across all the regions is the genetic modifying technique update. In 1990, there is no GM uptake usage across the counties, by 2019, most of the counties in regions 1,2,3,4 have achieved nearly 100% of GM uptake, but most of the counties in region 6 have less than 10% of GM uptake.

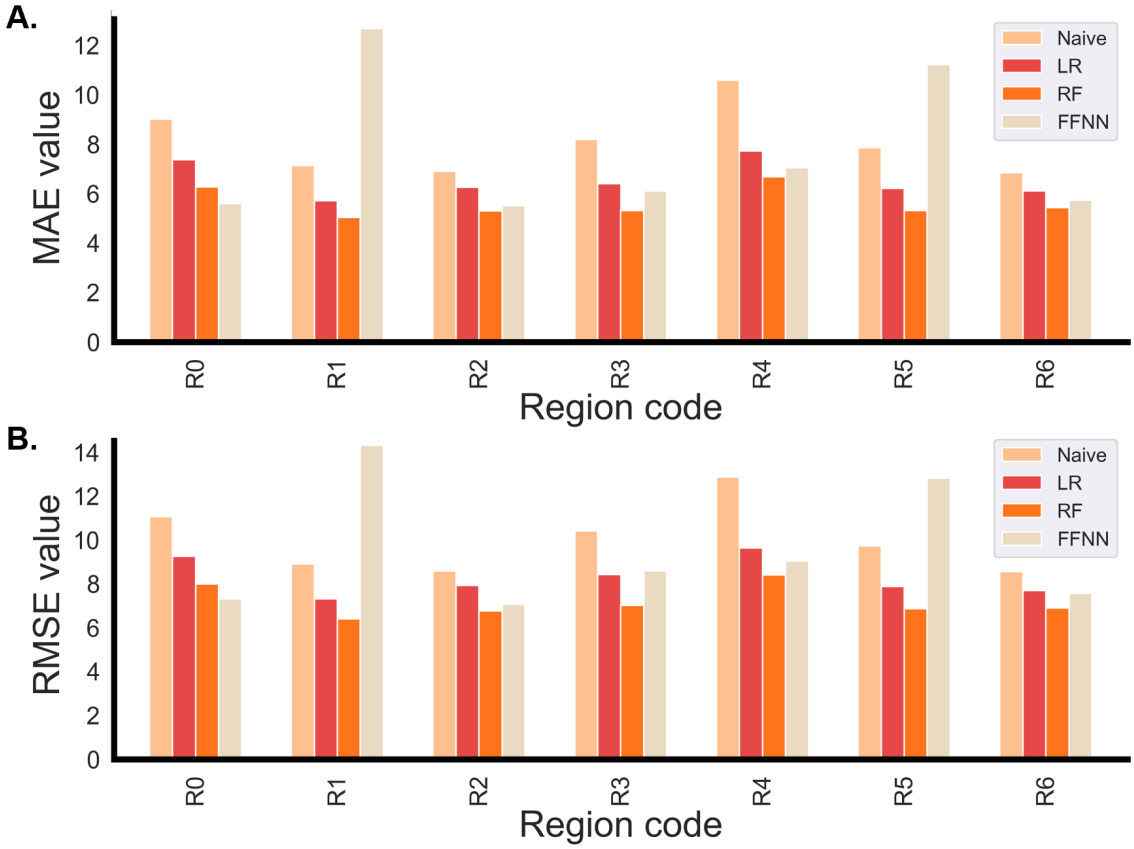

Figure 3: Models evaluation based on agricultural region code with MAE and RMSE, where R0 represents the integral dataset.

After data analysis, we noticed that agricultural regions can indeed have differences in soybean plantations. So we proposed that agricultural region-based estimation might be a better strategy for soybean yield prediction. After performing the classic statistical linear regression model (LR), the machine learning random forest estimator (RF), and the deep learning feed-forward neural network estimator (FFNN) on the integral dataset and the region-based dataset, we evaluated our model with the mean average error (MAE), root-mean-square deviation (RMSE), and proportion of the variance explained by the model (R^2) and get the model performance result.

The performance of different models was evaluated with the selected evaluation metrics and visualized with table and bar charts. In Table 1, we listed the model performance of Naive, linear regression, random forest, and FFNN under MAE, RMSE, and R^2 in seven different regions based on USDA agricultural regions.

Regarding MAE, a small value accounts for better model performance. Compared to the naive predic-

tion, both multiple linear regression and random forest regression models outstand naive prediction for all seven regions, indicating that both models are not under-fitting. Between these two models, random forest regression always performs better regardless of the region selected, which means the machine learning model surpassed the statistical model on this dataset. The feed-forward neural network model prevails over resting models in regard to the integral dataset; however, it is inferior to the random forest in R1 to R6, especially in R1 and R5 (Fig. 3). In addition, the table manifests that the linear regression and random forest model can generate stable results that are persistently better than the naive MAE, while FFNN is undulant, and sometimes it performs even worse than the naive MAE. This fluctuation in FFNN can be attributed to the different sizes of the dataset passed. Typically, neural network models require large amounts of data for training to generate accurate results. Hence, the FFNN model performs well on the integral dataset with numerous instances and when dealing with sub-datasets with fewer instances, deviated results were shown. The model performance evaluated with RMSE followed a similar trend to MAE evaluations. Same as MAE, the multiple linear regression and random forest regression model generates stable RMSE with values lower than the naive RMSE. The random forest regression model still performs better than linear regression for predicting soybean yield in all regions, including the entire U.S. (R0). For FFNN, we found FFNN generated unstable results and was underfitting for some situations. However, the FFNN model performs best when predicting soybean production using the whole dataset. In general, evaluated with MAE or RMSE, the random forest regression model turns out to be the most comprehensive in predicting the soybean yield per acre in each USDA agricultural region. If pursuing a quick insight into the soybean yield per acre in the entire U.S. with MAE, an FFNN model is recommended.

The model evaluation with R^2 can be interpreted as the proportion of the variation in the target features that is predictable from the other features. A larger R^2 value stands for better performance. Overall, the random forest regression model generated the best results, with steady and relatively high R^2 for each USDA agricultural region. FFNN model still had the best result for the prediction of soybean yield per acre with the whole dataset. Nevertheless, it is inferior to the random forest model on the small region-based dataset. For predicting R1, there is even a negative R^2 for the FFNN model. This unusual negative value can be attributed to the small dataset size and highly uncorrelated instances for FFNN to process. The model on R1 using FFNN is the only underfitted model.

Based on the results obtained from all three evaluation matrices (Fig. 3), the random forest regression model is the most effective model when considering predicting the soybean yield per acre based on the USDA agricultural region. The FFNN model is suggested when predicting the soybean yield per acre for all the soybean-cultivated regions in the United States. For the small region-based dataset, more neural network layers and smaller batch sizes should be employed to further verify its effectiveness. In addition, the FFNN models exhibited bad results in R1 regardless of the evaluation metrics used. One may consider exploring the potential underlying patterns in this region in the future. Besides, the machine learning and deep learning method outstand the statistical method in our prediction task.

5 Conclusion

Our project aims to find the most feasible strategy for soybean yield prediction. We performed exploratory data analysis, based on which we proposed that the region-based estimation would perform better (with higher accuracy) in soybean yield prediction. In order to find the most effective estimator for each region-based soybean dataset (including the whole region), we compared LR, RF, and FFNN models and evaluated the model performance with MAE, RMSE, and R^2, respectively.

In general, we noticed that FFNN had the best performance on the prediction of integral soybean cultivated region's yield, but its performance on the region-based dataset was unsatisfactory, especially for USDA agricultural region R1. However, the random forest regression model performed the best with the region-based dataset for all three evaluation metrics.

For future work, one may consider trying to find an independently collected test dataset to test the model performance. Besides, we would train our FFNN model with more layers and smaller batch sizes to increase our FFNN model learning ability, and RF with larger depth. Especially since we noticed a high error score for FFNN performance on region 1, we would perform a more thorough data analysis to find the potential characteristics of region 1 that attributing to the model performance. We can further analyze each model's

performance on their computational costs.

References

[1] Mian N Riaz. *Soy Applications in Food*. CRC Press, Boca Raton, FL, November 2005.

[2] Usda ers - soybeans and oil crops. https://www.ers.usda.gov/topics/crops/soybeans-and-oil-crops/. (Accessed on 12/04/2022).

[3] Mark Messina. Soy and health update: Evaluation of the clinical and epidemiologic literature. *Nutrients*, 8(12):754, November 2016.

[4] National oceanic and atmospheric administration. https://www.noaa.gov/. (Accessed on 12/04/2022).

[5] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[8] François Chollet et al. Keras. https://keras.io, 2015.