Crop prediction based on characteristics of the agriculture environment using various feature selection techniques and classifiers.

A. Sathvika Goud, Department of CSE, B.Tech student, CMR Technical Campus, Medchal road, Hyderabad, Telangana 501401
B. Nandini, Department of CSE, B.Tech student, CMR Technical Campus, Medchal road, Hyderabad, Telangana 501401
M. Supriya, Department of CSE, B.Tech student, CMR Technical Campus, Medchal road, Hyderabad, Telangana 501401
Ms. Tabeen Fatima, Assistant professor, CMR Technical Campus, Medchal road, Hyderabad, Telangana 501401

Abstract:

Agriculture is a growing field of research. In particular, crop prediction in agriculture is critical and is chiefly contingent upon soil and environment conditions, including rainfall, humidity, and temperature. In the past, farmers were able to decide on the crop to be cultivated, monitor its growth, and determine when it could be harvested. Today, however, rapid changes in environmental conditions have made it difficult for the farming community to continue to do so. Consequently, in recent years, machine learning techniques have taken over the task of prediction, and this work has used several of these to determine crop yield. To ensure that a given machine learning (ML) model works at a high level of precision, it is imperative to employ efficient feature selection methods to pre-process the raw data into an easily computable Machine Learning friendly dataset. To reduce redundancies and make the ML model more accurate, only data features that have a significant degree of relevance in determining the final output of the model must be employed. Thus, optimal feature selection arises to ensure that only the most relevant features are accepted as a part of the model. Conglomerating every single feature from raw data without checking for their role in the process of making the model will unnecessarily complicate our model. Furthermore, additional features which contribute little to the ML model will increase its time and space complexity and affect the accuracy of the model's output. The results depict that an ensemble technique offers better prediction accuracy than the existing classification technique.

Introduction

Crop prediction in agriculture is a complicated process and multiple models have been proposed and tested to this end. The problem calls for the use of assorted datasets, given that crop cultivation depends on biotic and abiotic factors. Biotic factors include those elements of the environment that occur as a result of the impact of living organisms (microorganisms, plants, animals, parasites, predators, pests), directly or indirectly, on other living organisms. This group also includes anthropogenic factors (fertilization, plant protection, irrigation, air pollution, water pollution and soils, etc.). These factors may contribute to the occurrence of many changes in the yield of crops, cause internal defects, shape defects and changes in the chemical composition of the plant yield. The shaping of the environment as well as the growth and quality of plants is influenced by abiotic and biotic factors. Abiotic factors can be divided into physical, chemical, and other. The recognized physical factors include: mechanical vibrations (vibration, noise), radiation (e.g., ionizing, electromagnetic, ultraviolet, infrared); climatic conditions (atmospheric pressure, temperature, humidity, air movements, sunlight); soil type, topography, soil rockiness, atmosphere, and water chemistry, especially salinity. The chemical factors include: priority environmental poisons, such as sulfur dioxide and derivatives, PAHs; nitrogen oxides and derivatives, fluorine, and its compounds, lead and its compounds, cadmium and its compounds, nitrogen fertilizers, pesticides, carbon monoxide. The others are: mercury, arsenic, dioxins and furans, asbestos, and aflatoxins. Abiotic factors also include bedrock, relief, climate, and water conditions - all of which affect its properties. Soil-forming factors have a diversified effect on the formation of soils and their agricultural value.

1.1 SOFTWARE REQUIREMENTS

Software requirements deal with defining software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or prerequisites are generally not included in the software installation package and need to be installed separately before the software is installed.

Platform – In computing, a platform describes some sort of framework, either in hardware or software, which allows software to run. Typical platforms include a computer's architecture, operating system, or programming languages and their runtime libraries.

Operating system is one of the first requirements mentioned when defining system requirements (software). Software may not be compatible with different versions of same line of operating systems, although some measure of backward compatibility is often maintained. For example, most software designed for Microsoft Windows XP does not run on Microsoft Windows 98, although the converse is not always true. Similarly, software designed using newer features of Linux Kernel v2.6 generally does not run or compile properly (or at all) on Linux distributions using Kernel v2.2 or v2.4.

APIs and drivers – Software making extensive use of special hardware devices, like high-end display adapters, needs special API or newer device drivers. A good example is DirectX, which is a collection of APIs for handling tasks related to multimedia, especially game programming, on Microsoft platforms.

Web browser – Most web applications and software depending heavily on Internet technologies make use of the default browser installed on system. Microsoft Internet Explorer is a frequent choice of software running on Microsoft Windows, which makes use of ActiveX controls, despite their vulnerabilities.

1)Visual Studio Community Version
2)Nodejs (Version 12.3.1)
3)Python IDEL (Python 3.7)

1.2 HARDWARE REQUIREMENTS

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware, A hardware requirements list is often accompanied by a hardware compatibility list (HCL), especially in case of operating systems. An HCL lists tested, compatible, and sometimes incompatible hardware devices for a particular operating system or application. The following sub-sections discuss the various aspects of hardware requirements.

Architecture – All computer operating systems are designed for a particular computer architecture. Most software applications are limited to particular operating systems running on particular architectures. Although architecture-independent operating systems and applications exist, most need to be recompiled to run on a new architecture. See also a list of common operating systems and their supporting architectures.

Processing power – The power of the central processing unit (CPU) is a fundamental system requirement for any software. Most software running on x86 architecture define processing power as the model and the clock speed of the CPU. Many other features of a CPU that influence its speed and power, like bus speed, cache, and MIPS are often ignored. This definition of power is often erroneous, as AMD Athlon and Intel Pentium CPUs at similar clock speed often have different throughput speeds. Intel Pentium CPUs have enjoyed a considerable degree of popularity, and are often mentioned in this category.

Memory – All software, when run, resides in the random access memory (RAM) of a computer. Memory requirements are defined after considering demands of the application, operating system, supporting software and files, and other running processes. Optimal performance of other unrelated software running on a multi-tasking computer system is also considered when defining this requirement.

Secondary storage – Hard-disk requirements vary, depending on the size of software installation, temporary files created and maintained while installing or running the software, and possible use of swap space (if RAM is insufficient).

Display adapter – Software requiring a better than average computer graphics display, like graphics editors and high-end games, often define high-end display adapters in the system requirements.

Peripherals – Some software applications need to make extensive and/or special use of some peripherals, demanding the higher performance or functionality of such peripherals. Such peripherals include CD-ROM drives, keyboards, pointing devices, network devices, etc.

2. LITERATURE SURVEY

2.1 Applying naive Bayes classification technique for classification of improved agricultural land soils:

https://www.researchgate.net/publication/309212171_Applying_Naive_Bayes_Data_Mining_Technique_for_Classification_of_Agricultural_Land_Soils

The advances in computing and information storage have provided vast amounts of data. The challenge has been to extract knowledge from this raw data that has lead to new methods and techniques such as data mining that can bridge the knowledge gap. This research aimed to assess these new data mining techniques and apply them to a soil science database to establish if meaningful relationships can be found. A large data set of Soil database is extracted from the Department of Soil Sciences and Agricultural Chemistry, S V Agricultural College, Tirupati, The database contains measurements of soil profile data from various locations of Chandragiri Mandal, Chittoor District. The research establishes whether Soils are Classified Using various data mining techniques. In addition, comparison was made between Naive bayes classification and analyse the most effective technique. The outcome of the research may have many benefits, t o agriculture, soil management and environmental.

2.2 Biotic components influencing the yield and quality of potato tubers

http://agronomyaustraliaproceedings.org/images/sampledata/2015_Conference/pdf/agronomy2015final00047.pdf

Potato yields in Canterbury have remained static at c. 60 t/ha for the last decade. In contrast, potato growth models predict potential yields of up to 90 t/ha, which have previously been achieved by some commercial growers. A two-year project conducted by industry and research partners has examined factors constraining crop yields. In year 1, 11 processing crops were intensively monitored (final yield, plant health and soil quality assessments) throughout the growing season. Soil-borne diseases (Rhizoctonia stem canker and Spongospora root infection) were identified as consistent factors in reduced yields, along with subsurface soil compaction and inadequate irrigation management. Cropping histories that included potatoes within the last 10 years resulted in faster onset of symptoms of Rhizoctonia stem canker (by emergence), compared with fields with periods of grass growth and no previous potato crops (8 weeks after emergence). In year 2, a controlled field experiment in a commercial crop (known to have high levels of soil-borne pathogens) attempted to isolate and quantify the impacts of soil-borne diseases on yield. Treatments included soil fumigant (90, 112 and 146 kg/ha chloropicrin), in-furrow application of azoxystrobin (1.5 l/ha) or flusulphamide (400 ml/ha), and a nil pesticide control. Soil-borne pathogen DNA tests before and after treatment showed a slight reduction in DNA levels of Rhizoctonia solani and Spongospora subterranea in the soil (plots treated with fumigant), but results were very variable. Final total fresh yield averaged 58 t/ha and did not differ between treatments. Throughout the season, the severity of R. solani on underground stems was consistently less for the azoxystrobin treatment compared to all other treatments.

2.3 Response surface methodology: A retrospective and literature survey

https://www.tandfonline.com/doi/abs/10.1080/00224065.2004.11980252

Response surface methodology (RSM) is a collection of statistical design and numerical optimization techniques used to optimize processes and product designs. The original work in this area dates from the 1950s and has been widely used, especially in the chemical and process industries. The last 15 years have seen the widespread application of RSM and many new developments. In this review paper we focus on RSM activities since 1989. We discuss current areas of research and mention some areas for future research.

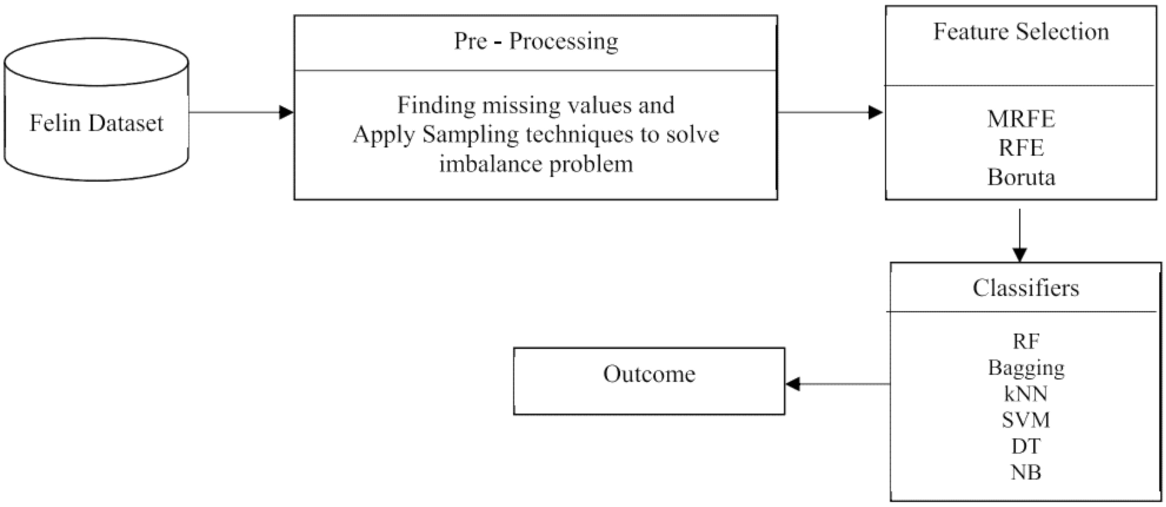
2.4 Application of response surface methodology for optimization of potato tuber yield

https://www.researchgate.net/publication/281612031_Application_of_Response_Surface_Methodology_for_Optimization_of_Potato_Tuber_Yield

The Author investigates the operating conditions required for optimal production of potato tuber yield in Kenya. This will help potato farmers to safe extra cost of input in potato farming. The potato production process was optimized by the application of factorial design 2 3 and response surface methodology. The combined effects of water, Nitrogen and Phosphorus mineral nutrients were investigated and optimized using response surface methodology. It was found that the optimum production conditions for the potato tuber yield were 70.04% irrigation water, 124.75Kg/Ha of Nitrogen supplied as urea and 191.04Kg/Ha phosphorus supplied as triple super phosphate. At the optimum condition one can reach to a potato tuber yield of 19.36Kg/plot of 1.8meters by 2.25 meters. Increased productivity of potatoes can improve the livelihood of smallholder potato farmers in Kenya and safe the farmers extra cost of input. Finally, i hope that the approach applied in this study of potatoes can be useful for research on other commodities, leading to a better understanding of overall crop production.

Proposed method

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.



1. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
2. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
3. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

MODULES:

1. importing the packages: using this module we will import all packages
2. exploring the dataset – Crop recommendation Data: Using this module we will upload dataset
3. data processing & cleaning: Using this module we will read data for processing
4. visualization using seaborn & matplotlib: Using this module will get graphical representation of information and data.
5. Splitting the data to train and test: Using this module will divide dataset into train & test for processing
6. building the model with and without feature selection
   - Feature Selection (SMOTE, ROSE, RFE, MRFE, BORUTA, MEMOTE)
   - KNN
   - Naive Bayes
   - Bagging Classifier
   - Random Forest
   - Decision Tree
   - SVM
   - Gradient Boosting
   - Voting Classifier
7. training the model: Using this module algorithms trained for processing & prediction building the model with Voting Classifier since it gives better accuracy comparing with Other Models
8. Flask Framework with Sq lite for signup and sign in: Using this module user will get register & login importing the packages

9. User gives input as Feature Values : Using this module user gives input for prediction the given input is preprocessed for prediction
10. trained model is used for prediction: Using this module predicted result displayed final outcome is displayed through frontend

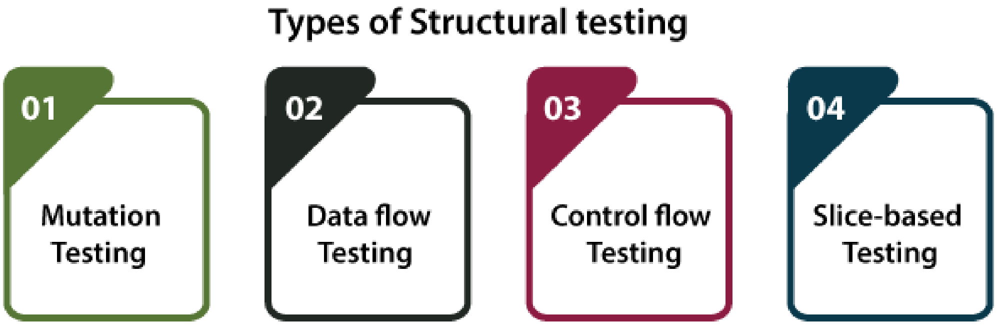
Note:
Extension - Voting Classifier, Gradient Boosting is build as Extension for feature values used for prediction since we it gives better accuracy around 100% comparing with other Models

ALGORITHMS:

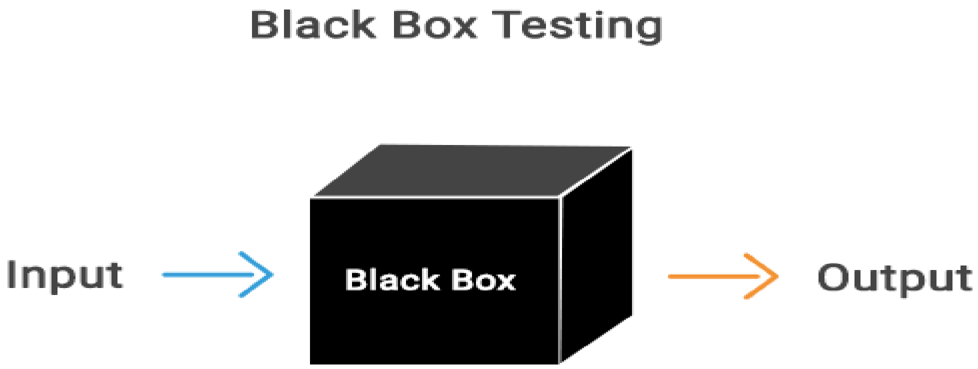
- Feature Selection (RFE, MRFE, BORUTA)
- **KNN:** The abbreviation KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.
- **Naive Bayes:** The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.
- **Bagging Classifier:** A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.
- **Random Forest:** Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- **Decision Tree:** Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable.
- **SVM:** Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- Gradient Boosting: Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

- Voting Classifier: A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output.

It is not possible to effectively test software without running it. Structural testing, also known as white-box testing, is required to detect and fix bugs and errors emerging during the pre-production stage of the software development process. At this stage, unit testing based on the software structure is performed using regression testing. In most cases, it is an automated process working within the test automation framework to speed up the development process at this stage. Developers and QA engineers have full access to the software's structure and data flows (data flows testing), so they could track any changes (mutation testing) in the system's behavior by comparing the tests' outcomes with the results of previous iterations (control flow testing).



Behavioral Testing:
The final stage of testing focuses on the software's reactions to various activities rather than on the mechanisms behind these reactions. In other words, behavioral testing, also known as black-box testing, presupposes running numerous tests, mostly manual, to see the product from the user's point of view. QA engineers usually have some specific information about a business or other purposes of the software ('the black box') to run usability tests, for example, and react to bugs as regular users of the product will do. Behavioral testing also may include automation (regression tests) to eliminate human error if repetitive activities are required. For example, you may need to fill 100 registration forms on the website to see how the product copes with such an activity, so the automation of this test is preferable.



TEST CASES:

S.NO	INPUT	If available	If not available
1	User signup	User get registered into the application	There is no process
2	User signin	User get login into the application	There is no process
3	Enter input for prediction	Prediction result displayed	There is no process

Conclusion

Predicting crops for cultivation in agriculture is a difficult task. This paper has used a range of feature selection and classification techniques to predict yield size of plant cultivations. The results depict that an ensemble technique offers better prediction accuracy than the existing classification technique. Forecasting the area of cereals, potatoes and other energy crops can be used to plan the structure of their sowing, both on the farm and country scale. The use of modern forecasting techniques can bring measurable financial benefits.

References

[1] R. Jahan, ‘‘Applying naive Bayes classification technique for classification of improved agricultural land soils,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.

[2] B. B. Sawicka and B. Krochmal-Marczak, ‘‘Biotic components influencing the yield and quality of potato tubers,’’ Herbalism, vol. 1, no. 3, pp. 125–136, 2017.

[3] B. Sawicka, A. H. Noaema, and A. GÆowacka, ‘‘The predicting the size of the potato acreage as a raw material for bioethanol production,’’ in Alternative Energy Sources, B. Zdunek, M. OlszÆwka, Eds. Lublin, Poland: Wydawnictwo Naukowe TYGIEL, 2016, pp. 158–172.

[4] B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, ‘‘Biotic and abiotic factors influencing on the environment and growth of plants,’’ (in Polish), in Proc. Bioró»norodno–¢ Środowiska Znaczenie, Problemy, Wyzwania. Materia“y Konferencyjne, Pu“awy, May 2017. [Online]. Available: https://bookcrossing.pl/ksiazka/321192

[5] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ‘‘Response surface methodology: A retrospective and literature survey,’’ J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2004.

[6] D. K. Muriithi, ‘‘Application of response surface methodology for optimization of potato tuber yield,’’ Amer. J. Theor. Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[7] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ‘‘Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,’’ Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183–188, 2014.

[8] J. R. Olƒdzki, ‘‘The report on the state of remotesensing in Poland in 2011–2014,’’ (in Polish), Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2015.

[9] K. Grabowska, A. Dymerska, K. PoÆarska, and J. Grabowski, ‘‘Predicting of blue lupine yields based on the selected climate change scenarios,’’ Acta Agroph., vol. 23, no. 3, pp. 363–380, 2016.

[10] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ‘‘Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning,’’ Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.

[11] N. Chanamarn, K. Tamee, and P. Sittidech, ‘‘Stacking technique for academic achievement prediction,’’ in Proc. Int. Workshop Smart Info-Media Syst., 2016, pp. 14–17.

[12] W. Paja, K. Pancerz, and P. Grochowalski, ‘‘Generational feature elimination and some other ranking feature selection methods,’’ in Advances in Feature Selection for Data and Pattern Recognition, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97–112.

[13] D. C. Duro, S. E. Franklin, and M. G. DubØ, ‘‘A comparison of pixelbased and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery,’’ Remote Sens. Environ., vol. 118, pp. 259–272, Mar. 2012.

[14] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, ‘‘Soil classification and suitable crop prediction,’’ in Proc. Nat. Conf. Comput. Biol., Commun., Data Anal. 2017, pp. 25–29.

[15] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, ‘‘Deep Gaussian process for crop yield prediction based on remote sensing data,’’ in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1, pp. 4559–4565.